UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE INFORMÁTICA
DEPARTAMENTO DE ARQUITECTURA DE COMPUTADORES Y AUTOMÁTICA

TESIS DOCTORAL

Herramientas eficientes para el análisis masivo de datos ómicos

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Daniel Tabas Madrid

DIRECTORES

Alberto Pascual Montano
Carlos García Sánchez

Madrid, 2018

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE INFORMÁTICA

Departamento de Arquitectura de Computadores y Automática

TESIS DOCTORAL

HERRAMIENTAS EFICIENTES PARA EL ANÁLISIS MASIVO DE DATOS ÓMICOS

MEMORIA PARA OPTAR AL GRADO DE DOCTOR
PRESENTADA POR:

Daniel Tabas Madrid

DIRECTORES:

Alberto Pascual Montano
Carlos García Sánchez

Madrid, 2017

A mi abuelo Matías, estarías orgulloso.
A mis padres, a mi hermana y a María.

Agradecimientos

Quiero empezar dando las gracias a mi director, Alberto, me enseñaste la bioinformática, primero como profesor, y después como jefe. Eres un modelo a seguir, pones pasión y esfuerzo en lo que haces. También has depositado muchísima confianza en mí, que al cabo de unos años se ha transformado en la tesis que termino ahora.

Quiero dar también las gracias a mi tutor, Carlos, por acompañarme desde hace ya bastantes años, desde mi época de administrador en *ArTeCS*, ser tan cercano y haber aceptado ser partícipe de este reto.

A mis padres, por los cuales soy quien soy, que siempre han confiado en mí y me han apoyado y aconsejado en todos los pasos que he dado en la vida. También a mi hermana Rosa, que siempre ha estado cerca y me ha apoyado también. Os quiero mucho.

A María, que ha sido la que más ha sufrido esta tesis, gracias por aguantar mis momentos de bajón, mi mal humor, y gracias también por ayudarme tanto en todo y darme la confianza que necesitaba, sin tu apoyo esto no habría sido posible.

Haciendo memoria, quiero pensar que toda esta aventura empezó antes de terminar la carrera, con Christian y Nacho y el proyecto de fin de carrera que me ofrecieron, los cuales me animaron a seguir después en la universidad y estudiar el Máster de investigación. Todo esto se materializó en poder empezar a trabajar en *ArTeCS* y compartir muchos momentos con mucha gente buena: Rober, Ezequiel, Luis Canet, Marco, Jorge, Rodri, Javi Setoain, Carolina, Darío, Hugo, Fermín, Silvio, Ricardo, Antón, Edgardo, Juan Carlos, Carlos Juega, Luis Piñuel, Manuel Prieto, Guillermo, David y todo el resto de la gente del departamento.

Mención especial también para Rubén, por el que pude empezar a trabajar en el *CNB* con Alberto, y que acabamos compartiendo birras y zumos de tomate. A toda la gente con la que he compartido momentos en el CNB: a los proteómicos (Salva, Migue, Vital, Alberto Paradela, Adán, Sergio, Carmen y demás gente del laboratorio, y un recuerdo especial para Juan Pablo Albar, gracias al cual pudimos introducir la patita en la proteómica, y cuyo fallecimiento nos entristeció mucho a todos), a Carlos Alonso, a Oliveros, a toda la gente de *TIC* (en especial a Íñigo y Sonia), a toda la gente del B13: Joan, Rubén, Kino, Javi Vargas, Melero, Jesús Cuenca, Laura, Nacho, Jordi, Josemi, Josué, Vero, Johan, Pablo, Juan, Vilas, COSS, y tanta otra gente que ha pasado por el labo, y muy especialmente a José María Carazo, que siempre me ha ofrecido un hueco con ellos, y a Blanca, que me ha ofrecido su ayuda en todo, y con la que más comidas en la cafetería he compartido. También a Ilaria, y en especial a todos los compis del M3, con los que más momentos compartí: Rubén, que ya te mencioné antes; Danny, que te convertiste en un hermano para mí; Rafa; João, con el que pasamos muy buenos momentos; Mònica, que ya sabes que somos muy distintos, pero aprendimos a trabajar juntos, y nos apoyamos mucho en la marcha de Alberto, de verdad que te echo mucho de menos; Javi Setoain, qué bien volver a coincidir; y Marta. En definitiva, guardo muy buen recuerdo de toda la gente con la que he coincidido allí.

Agradecer también a la gente de otros centros con la que he compartido congresos y con la que he colaborado en mi época en el CNB, Javier de la Rivas, Ander, Ángel Rubio, Fernando Corrales, Víctor Segura y Elizabeth Guruceaga, Gorka, etc, ha sido muy enriquecedor poder compartir momentos, ideas y opiniones.

Gracias a la gente de *Perkinelmer*, en especial a los ex-*Integromics* (Eduardo, David, Vinicio, Juan, Migue, Simone), por haberme acogido tan bien en tan poco tiempo.

Quiero dar las gracias también a toda la gente que tengo cerca y que me han apoyado, a toda mi familia, a la familia de María que me ha aceptado como uno más, y a mis amigos. A Javi y Miriam, que nos queremos y tratamos como primos y a todo el grupo (Luis, Elisa, Juan, Maika y Gabi) con el que poder desconectar un rato los fines de semana. A Pichu, Yiropa y Pke, por todas las buenas conversaciones y el *hate* amistoso. Espero seguir siendo el pegamento, y cuidado que ahora tendré más tiempo para entrenar y os voy a quitar los *KOMs* en el *Strava*. A Jorge y a Rober, que ya os mencioné antes, y que aunque no nos veamos mucho, tengo muy buenos recuerdos de todo lo que vivimos juntos en la uni y en el curro. A Sergio, que podemos hablar de cualquier cosa, eres un gran ingeniero y un gran fotógrafo, pero sobre todo, eres un

gran tío. A Héctor, por ofrecerme otra perspectiva de la vida, seguro que un día de estos sale una foto tuya en la *Vice*. Al resto de amigos de la facultad, en especial a Julio César, Cris, Emi y demás gente a la que quiero un montón, y que nos vemos menos de lo que deberíamos. Y a todos los que no he mencionado, pero que estáis o habéis estado ahí, soy lo que soy en parte gracias a todos vosotros.

Índice general

Índice de figuras

Índice de tablas

Resumen

En los últimos años se han desarrollado técnicas en el campo de la biología que han revolucionado las áreas de la genómica y la proteómica. Estas técnicas, entre las que se encuentran la secuenciación masiva y la proteómica de *Shotgun*, nos están permitiendo un conocimiento mucho más profundo del funcionamiento de las células, pudiendo ver qué ARN mensajero y proteínas están presentes en un momento puntual de las mismas, además de conocer mejor algunos mecanismos de regulación. Con el desarrollo de estas tecnologías, se están generando más datos de los que es posible procesar en una cantidad razonable de tiempo. Es necesario el desarrollo de nuevas herramientas que manejen este tipo de datos de una forma eficiente, haciendo uso de técnicas de computación de altas prestaciones que incluyan el uso de granjas de computación, computación paralela y gestión de plataformas virtualizadas. En la presente tesis se pretende realizar un abordaje integral del análisis masivo de datos provenientes de estas técnicas con herramientas eficientes, empezando por el procesamiento de los datos en crudo y obteniendo información de más alto nivel sobre expresión de genes y proteínas, enriqueciéndola con información relacionada de bases de datos y ontologías de libre acceso, para finalmente generar informes que reflejen el funcionamiento celular asociado a toda esta información. También incluye el desarrollo de herramientas generadoras de hipótesis en el ámbito de la regulación génica, que sirvan a biólogos experimentalistas para el desarrollo de nuevos experimentos de validación.

Este abordaje se ha concretado en el desarrollo de diferentes metodologías y herramientas. Primeramente se han desarrollado varios flujos de trabajo para análisis de *RNA-Seq*, *Microarrays* y proteómica de *Shotgun* de diferentes proyectos y bases de datos públicas tales como *ENCODE*, *Human Proteome Project*, *Illumina Human Body Map* o *the Cancer Cell Line Encyclopedia*, enfocados para realizar estudios proteogenómicos, permitiendo detectar con exactitud los genes expresados sin necesidad de un control, o mezclar datos transcriptómicos

y proteómicos para poder realizar una mejor detección de proteínas. Los resultados de estos flujos aplicados a los datos de los diferentes proyectos mencionados se han recogido en un panel web que permite su búsqueda y visualización interactiva (`http://sphppdashboard.cnb.csic.es/`). Se ha desarrollado también una nueva versión de una herramienta web de enriquecimiento modular de listas de genes, permitiendo su utilización con una gran cantidad de organismos y bases de datos de anotaciones, habiendo incluido en esta versión la posibilidad de comparar dos listas de genes, y habiendo mejorado además la eficiencia de la herramienta y la visualización de resultados (`http://genecodis.cnb.csic.es/`). Se incluye además el desarrollo una nueva metodología de predicción de interacciones entre micro ARNs y mensajeros, basada en la combinación de información de bases de datos de predicciones existentes, calculando su precisión en base a las interacciones ya validadas y generando una nueva puntuación relacionada con la probabilidad de que una predicción pueda existir realmente. A partir de esta metodología se ha desarrollado además una herramienta web para poder acceder a la base de datos de interacciones generada a través de la misma (`http://m3rna.cnb.csic.es/`). Otro de los métodos desarrollados consiste en un comparador de perfiles de expresión génica, con una estadística de comparación de rangos, que permite otorgar una puntuación a cada comparación. Este método se ha utilizado para comparar un perfil de expresión de entrada con los extraídos de diferentes bases de datos públicas como *GEO*, *Connectivity Map* y *DrugMatrix*, que han sido a su vez procesados y asociados con información de fármacos y enfermedades, que permiten finalmente hacer nuevas inferencias en el contexto del reposicionamiento de fármacos. A partir de esta metodología, y con la base de datos generada a partir del procesamiento de las tres bases de datos mencionadas, se ha desarrollado una herramienta web para poder realizar estas comparaciones (`http://nffinder.cnb.csic.es/`). Por último, se ha creado una herramienta web que permite integrar información estructural y de secuencia de proteínas a partir de la información procedente de *EMDB*, *PDB* y *Uniprot*, pudiendo visualizar a nivel de estructura anotaciones de secuencias procedentes de diversas bases de datos públicas como *dSysMap*, *bioMuta* o *PhosphoSitePlus* (`http://3dbionotes.cnb.csic.es/`).

Abstract

In recent years, techniques in the field of biology that have revolutionized the areas of genomics and proteomics have been developed. These techniques, including high throughput sequencing and Shotgun proteomics, are allowing us a much deeper understanding of the cells' behavior, being able to see which messenger RNA and proteins are present on a certain moment, also allowing to know better some mechanisms of regulation. With the development of these technologies, more data than is possible to process in a reasonable amount of time is being generated. It is necessary to develop new tools that handle this type of data in an efficient way, making use of high performance computing techniques that include the use of computer clusters, parallel computing and management of virtualized platforms. The intention of this work is to carry out an integral approach to the analysis of the data coming from these techniques with efficient tools, starting with the processing of raw data and obtaining high level information on gene and protein expression, enriching it with related information of ontologies and free access databases in ordert to create reports that reflect the cellular behavior associated with all that information. It also includes the development of hypothesis-generating tools in the field of genetic regulation, which allows experimental biologists the development of new validation experiments.

This approach has resulted in the development of different methodologies and tools. First, several workflows for the analysis of *RNA-Seq*, Microarrays and Shotgun proteomics of different projects and public databases such as *ENCODE*, *Human Proteome Project*, *Illumina Human Body Map* or *Encyclopedia of Cancer Cell Line* have been developed, focused on performing proteogenomic studies, allowing an accurate detection of expressed genes without the need of controls, or mixing transcriptomic and proteomic data to enable better protein detection. The results of these workflows applied to the data coming from the different projects have been collected in a web panel that allow their search and interactive visualiza-

tion (http://sphppdashboard.cnb.csic.es/). A new version of a modular enrichment tool of gene lists has also been developed, allowing its use with a large number of organisms and annotation databases, and which has included in this version the possibility of comparing two lists of genes, having also improved the efficiency of the tool and visualization of results (http://genecodis.cnb.csic.es/). This work also includes the development of a new methodology for predicting interactions between microRNAs and messenger RNA, based on the combination of information from existing prediction databases, calculating their accuracy based on previously validated interactions, and generating a new score related to the probability that a prediction is actually possible. From this methodology, a web tool has been developed in order to access the database of interactions generated through this new method (http://m3rna.cnb.csic.es/). Another of the methods developed consists on a comparator of gene expression profiles, with a statistic based on comparison of ranges, that assigns a score to each comparison. This method has been used to compare an input expression profile with those extracted from different public databases such as *GEO*, *Connectivity Map* and *DrugMatrix*, which have in turn been processed and associated with drug and disease information, which finally allows to create new inferences in the context of drug repositioning. From this methodology, and with a database generated from the processing of the three mentioned repositories, a tool to perform these comparisons has been developed (http://nffinder.cnb.csic.es/). Finally, a tool that allows the integration of protein structural and sequence information using data coming from *EMDB*, *PDB* and *Uniprot* has been created, allowing the visualization at structure level of sequence annotations from various public databases such as *dSysMap*, *bioMuta* or *PhosphoSitePlus* (http://3dbionotes.cnb.csic.es/).

Capítulo 1

Introducción

1.1. Ómicas

Ómica es una palabra proveniente del inglés que se refiere al estudio de un conjunto en su totalidad. En el área de la biología molecular se utiliza para definir el estudio de los diferentes sistemas biológicos que conforman el funcionamiento de las células[221]. Varias de estas ómicas, tales como la genómica, la transcriptómica, la proteómica o la metabolómica entre otros, estudian conjuntamente los procesos de expresión de los genes y sus productos, y podemos englobarlas con el nombre genérico de genómica funcional[112] tal y como se representa en la Figura 1.1.

Desde hace varias décadas se ha intentado abordar el estudio de estos sistemas biológicos, pero el auge de estas ómicas ha venido dado por el desarrollo de técnicas de alto rendimiento en el área de la biología, las cuales son capaces de generar grandes volúmenes de datos[180] que contienen información suficiente del sistema como para estudiarlo. Anteriormente al desarrollo de estas técnicas, el estudio completo de un sistema biológico era una tarea casi imposible. El uso de la bioinformática era mucho más restringido, puesto que normalmente los estudios se limitaban a un elemento o interacción en concreto, como por ejemplo averiguar la existencia de una reacción química, o saber si un gen se expresaba o no. En la actualidad estas nuevas técnicas son capaces de analizar procesos o sistemas completos correspondientes a las células, generando como se ha dicho anteriormente cantidades ingentes de datos.

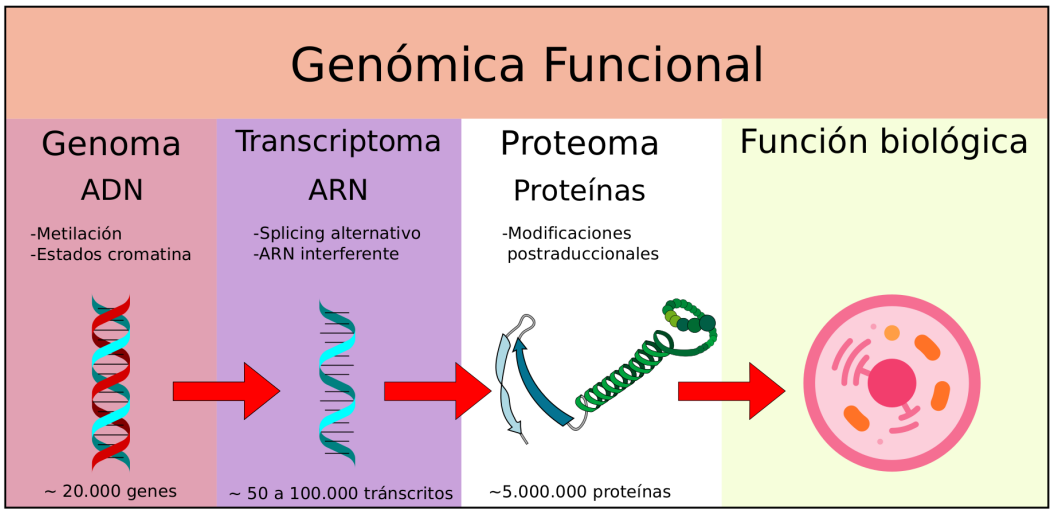


Figura 1.1: Áreas de la biología estudiadas por la genómica funcional.

De hecho, el crecimiento de la cantidad de datos biológicos y la disminución del coste de su generación están siendo mucho mayores que las previsiones[19, 203], generándose dos problemas, uno de almacenamiento, y otro de procesamiento. El problema de almacenamiento viene causado porque la generación de los datos está siendo más rápida que el desarrollo de tecnologías para su procesamiento y almacenamiento. Además, el desarrollo de nuevos y mejores sistemas de almacenamiento también es más lento que la evolución de estas técnicas biológicas, provocando debido a esta brecha un cada vez mayor coste en la adquisición de los mismos. En este sentido una de las prioridades es la de generar flujos de trabajo para procesar los datos lo antes posible, y solamente almacenar los resultados de este procesamiento, que contienen sólo la parte necesaria de los datos originales, y tienen un menor tamaño. El problema del procesamiento consiste en que el crecimiento de estos datos hace que los algoritmos que se utilizaban tradicionalmente para analizar este tipo de información sean ineficientes debido a su gran volumen, siendo imposible utilizarlos en un tiempo razonable. Además, las técnicas estadísticas para procesar grandes conjuntos de datos necesitan ser adaptadas a esta nueva realidad. Las soluciones en este sentido se centran por un lado en construir representaciones de los datos muy eficientes a la hora de ser mapeados en memoria, y por otro lado en utilizar todos los recursos posibles de computación, sobre todo mediante el uso de paralelismo.

En este punto la bioinformática, definida como la aplicación de técnicas computacionales en el ámbito de datos biológicos, cobra una gran importancia[115]. La heterogeneidad de los problemas a resolver y el formato de los datos presentes en bioinformática es bastante grande,

incluso dentro de un mismo área de conocimiento, por lo que no existe una metodología genérica. Lo más correcto es definir formatos de representación de los datos y flujos de trabajo para cada una de las áreas de análisis biológicos por separado, que como hemos dicho anteriormente conforman las ómicas.

El trabajo presentado en este manuscrito está centrado en la creación de herramientas y metodologías dentro del área de la genómica funcional. De las ómicas presentes en este área nos hemos centrado sobre todo en la transcriptómica y la proteómica.

1.1.1. Computación de alto rendimiento en ómicas

Una de las formas de afrontar este tan rápido crecimiento en la generación de datos biológicos es la utilización de sistemas de computación de alto rendimiento, para que de esta forma puedan ser procesados de forma más rápida y eficiente. Hasta la pasada década, el desarrollo de los procesadores estuvo guiado por el incremento del rendimiento mediante la explotación del paralelismo a nivel de instrucción y el escalado de la frecuencia de reloj como consecuencia de una mayor capacidad de integración en los chips conocida como Ley de Moore[200]. Se desarrollaron innumerables técnicas para mejorar el rendimiento en este tipo de arquitecturas, tales como: la ejecución fuera de orden, que permitía aprovechar ciclos de reloj reorganizando el orden de las instrucciones teniendo en cuenta las dependencias de datos; la ejecución especulativa[266], que a través de la predicción de saltos permitía ejecutar instrucciones posteriores a un salto condicional, ahorrando ciclos de reloj en caso de éxito; el *Single Instruction, Multiple Data* (*SIMD*), incluido en procesadores *Intel* como extensiones *SSE* y en *AMD* como *3DNow!*, y que permitía realizar la misma operación sobre un conjunto de datos en paralelo; el uso de memorias caché más grande y con más niveles, permitiendo reducir las diferencias entre el procesador y la memoria principal en lo que se vino a llamar el *memory wall*[311]; o el *Simultaneous Multithreading*[286] en el que gracias a la duplicación de varios elementos del procesador, éste podía permitir varios hilos de ejecución de forma simultánea. A día de hoy estas mejoras y sus evoluciones están presentes en los procesadores de propósitos general, destacando el crecimiento en el ancho de vector de hasta 512-bits para las extensiones multimedia en el caso de las AVX-512[268] de *Intel*.

Pero llegó un momento en el que el incremento de prestaciones que los procesadores solían incorporar en un chip llegó a un límite. La frecuencia de reloj no pudo seguir aumentando[225]

a causa de elevados consumos energéticos lo que suponía una gran dificultad a la hora de disipar la cantidad de calor producida en el chip. De manera análoga a problemas como el anteriormente citado *memory wall* o el *ILP wall*, que es el límite de instrucciones simultáneas que el procesador es capaz de ejecutar, irrumpe con fuerza el conocido como *power wall* que se acentúa con el incremento de la frecuencia de reloj y la escala de integración. Este aspecto ha supuesto un cambio radical a la hora de diseñar microprocesadores en la actualidad, poniendo el foco en la explotación de tareas entre procesadores con la aparición de los procesadores con varios núcleos, o *multi-core*, que consisten en la integración de varios procesadores independientes en el mismo espacio del chip. Estos procesadores destacan por ser más sencillos y por ende menos exigentes desde el punto de vista del consumo. Al existir varios núcleos, se podían seguir varios flujos de procesamiento de forma independiente y paralela, y al ser éstos más sencillos, el consumo energético se redujo considerablemente. Este nuevo paradigma trajo consigo otra dificultad añadida, la necesidad de que tanto los sistemas operativos como las aplicaciones explotaran explícitamente este nuevo paradigma de paralelismo con el fin de poder aprovechar toda su potencia computacional. Este tipo de procesadores ha seguido evolucionado durante esta década, a mayor escala de integración se desarrollan chips con mayor número de núcleos con caches más grandes.

En tareas en las que se procesan cantidades ingentes de datos como puede ser en el contexto bioinformático, el uso de un único procesador no suele ser la solución más adecuada en vista del bajo rendimiento observado, por lo que la tendencia es a hacer uso de computadores mayores con varios chips *multi-core* que cooperaran para realizar tareas de forma más eficiente. En un inicio se desarrollaron equipos especiales, como *mainframes*, que tenían propósitos muy específicos, y costes muy elevados. En el área de la bioinformática, por ejemplo, existieron potentes servidores especializados en realizar tareas tales como alineamientos de secuencias contra bases de datos, como *BLAST*[9]. Posteriormente se exploraron los *clusters*, que son conjuntos de ordenadores unidos entre sí por una red, vistos de forma externa como un único recurso computacional. Los utilización de *clusters* viene motivada por la posibilidad de utilizar sistemas más baratos, y que por su flexibilidad, permitiesen añadir máquinas con diferente configuración *hardware*, e incluso diferentes sistemas operativos.

Los *clusters* de computación siguen siendo la alternativa más utilizada hoy en día para llevar a cabo tareas complejas en el contexto bioinformático. Sin embargo, con el auge de los aceleradores tipo *GPUs (Graphics Processors Units)* con sorprendentes rendimientos, la comunidad científica ha mostrado interés en la migración de aplicaciones biológicas que presenten un alto

grado de paralelismo a este tipo de arquitecturas. Es tal el interés creciente, que a modo de ejemplo en el contexto de alineamiento de secuencias mediante el algoritmo de *Smith-Waterman* para proteínas hemos observado la aparición de varios desarrollos destacando el *CUDA-SW++*[165] y sus versiones sucesivas *CUDA-SW++2.0*[168] y *CUDA-SW++3.0*[166].

Sin embargo, el auge de los aceleradores en el contexto *HPC (High Performance Computing)* no se restringe únicamente al ámbito de las *GPU* donde *NVIDIA* aparece en posición dominante con el framework *CUDA*[212]. En los últimos años destacamos la aparición del coprocesador *Xeon-Phi*. *Intel* comercializa la primera versión de la arquitectura *MIC (Many Integrated Cores)* en el año 2012 bajo el nombre de *Knights-Corner (KNC)* que deriva del difunto proyecto *Larrabee*. Esta arquitectura se basa en dotar al chip de un conjunto elevado de *cores* con arquitectura similar a los *Pentium* a los que se les incorporan potentes unidades vectoriales de 512-bits, todo ello interconectado con un anillo bidireccional. Desde el año pasado se comercializa la segunda versión de la arquitectura *MIC* bajo el nombre de *Knights-Landing (KNL)* cuyas principales mejoras redundan en incrementar las capacidades de las unidades vectoriales unificando su repertorio con los procesadores de propósito general *(AVX-512 SIMD)*, mayor ancho de banda con memoria y la posibilidad de configuración de dicha memoria *HBM (High-Bandwidth Memory) ad-hoc* según la aplicación. Entre las principales ventajas de la arquitectura *MIC* frente a la arquitectura *GPU* de *NVIDIA* podemos destacar su facilidad de programación debido a que los modelos de programación soportados son idénticos a los utilizados en los procesadores de propósito general, por lo que la tarea de migración de código se ve claramente beneficiada, pudiendo incluso reutilizar el código y binario desarrollado para arquitecturas *Intel-x86*. De manera análoga a lo observado con el uso de las *GPUs*, destacamos el interés de la comunidad por migrar herramientas a estas arquitecturas más eficientes desde el punto de vista computacional. Siguiendo el símil anterior, en el contexto de alineamiento de secuencias mediante el algoritmo *Smith-Waterman* destacamos para el *KNC* las herramientas *SWAPHI*[167], *Parasail*[56] y *SWIMM*[246].

En el contexto de fabricantes de procesadores también es importante destacar el movimiento relativamente reciente por parte de *Intel* en la adquisición de uno de los principales fabricantes de *FPGA (Field-Programmable Gate Array)* en el año 2015[1]. *Intel* adquiere *Altera* con el fin de dotar a sus procesadores de alta gama de *hardware* reconfigurable que permita incrementar el rendimiento principalmente el aplicaciones del ámbito transaccional. Este aspecto queda patente con el anuncio de Microsoft de hacer uso de estas capacidades de las *FPGAs* en su

[1] `https://www.nytimes.com/2015/06/02/business/dealbook/intel-altera-buy-chips-computers-chip-maker.html`

infraestructura de *Cloud Azure*[2] y con el acuerdo de *Amazon* con el otro fabricante destacado de *FPGAs Xilinx* para dotar a sus servicios *Amazon Web Services* de capacidades computacionales basadas en *FPGA*. Recientemente *Intel* ha anunciando también un acuerdo con el gigante de venta por Internet *Alibaba* para dotar a su *Cloud* de recursos computacionales basados en sus *FPGA*[3]. Ya existen trabajos que ponen de relevancia las ventajas computacionales de este tipo de aceleradores en el contexto de alineamiento de secuencias destacando la herramienta *OSWALD*[247] para las *FPGAs* de *Intel-Altera* que puede lograr rendimientos parecidos a las *GPUs*.

En el contexto de la computación en la nube o *Cloud Computing* conviene destacar el fenómeno de utilización creciente de este tipo de infraestructura por parte de algunos centros bioinformáticos. La utilización de recursos computacionales en la nube viene motivado principalmente por el abaratamiento de precios de los principales proveedores de servicio como *Amazon AWS, IBM Cloud, Microsoft Azure* o *Google Cloud*. La facilidad a la hora de desplegar las herramientas informáticas, el pago por uso en contraposición de la amortización de una infraestructura informática fija y las ventajas en el mantenimiento de los equipos han motivado a algunos centros de investigación a plantearse la computación en la nube como alternativa. Estos aspectos van de la mano del desarrollo más eficiente de infraestructura *software* que facilite la distribución de la información entre los diferentes nodos y su computación en paralelo. En este ámbito destacamos la aparición del modelo de programación distribuida *Map-Reduce* y su posterior popularización con el *framework Hadoop* desarrollado por *Google* que permite trabajar con un alto número de nodos y volúmenes de datos de manera sencilla. En la actualidad destacamos el *framework Spark* como evolución de *Hadoop* siendo este más eficiente en el procesamiento de información al reducir significativamente la comunicación entre nodos y gestionar eficientemente la información en la memoria de los nodos. Como alternativa que va ganando mayor aceptación entre la comunidad de investigadores bioinformáticos destacamos la *API Spark Python* que une dos mundos en claro avance: la computación distribuida por medio de *Spark* y el incremento de desarrolladores de *Python*[4] que va desbancando al popular lenguaje *R* ampliamente utilizado en el campo del *Big Data*.

[2] https://www.top500.org/news/microsoft-goes-all-in-for-fpgas-to-build-out-cloud-based-ai/

[3] https://newsroom.intel.com/news/alibaba-collaborating-intel-fpga-based-solution-help-customers-accelerate-business-applications

[4] http://spectrum.ieee.org/computing/software/the-2016-top-programming-languages

1.1.2. Transcriptómica

1.1.2.1. Historia

El ADN de las células se transcribe en ARN gracias a la acción de la enzima ARN polimerasa, incluyendo varios tipos como el ARN mensajero, el ARN de transferencia, el ARN ribosomal, el ARN no codificante o el ARN pequeño entre otros. Este proceso, llamado transcripción del ADN, forma parte de la maquinaria de expresión génica de las células, transmitiendo información que será codificada en proteínas, o en forma de elementos regulatorios[6]. Los mecanismos de expresión génica son diferentes en organismos eucariotas y procariotas[271]. Centrándonos en eucariotas, el proceso se realiza en el núcleo de sus células, ocurriendo después de la transcripción propiamente dicha un proceso de maduración del ARN, que permite generar a partir de un tránscrito primario distintas moléculas de ARN maduro. Esta maduración viene dada por el *splicing* alternativo, que elimina los intrones del ARN primario y además puede eliminar algunos de los exones del mismo[199].

La transcriptómica es el área que estudia la expresión de los genes mediante el estudio del ARN, apareciendo esta palabra por primera vez en 1996[230, 295]. La caracterización y clasificación del ARN, así como el estudio de la estructura transcripcional de los genes mediante la determinación de sus extremos, patrones de *splicing* y otras modificaciones posteriores a la transcripción, así como la cuantificación de los tránscritos y su comparación en diferentes condiciones de la célula son los trabajos más importantes que se realizan dentro de este área.

A pesar de la cantidad de tipos de ARN que forman el transcriptoma, en muchas ocasiones la transcriptómica se centra en el análisis de abundancia del ARN mensajero maduro, lo que se conoce como la realización de perfiles de expresión génica. Con este tipo de análisis se genera información acerca de los genes expresados y sus niveles de expresión, con la posibilidad también de extraer información acerca de las diferentes isoformas expresadas. En la actualidad existen técnicas de alto rendimiento como los chips de ADN (*Microarrays*) y la ultra-secuenciación que permiten realizar este tipo de análisis de forma rápida y eficaz, lo que ha llevado a una revolución en el área[279].

1.1.2.1.1 *Microchips de ADN (Microarrays)*

La primera técnica de alto rendimiento usada en biología que se desarrolló fue la de los chips de ADN (*DNA Microarrays*)[41, 116, 285]. Estos chips consisten en una superficie sólida en la cual se deposita una colección de fragmentos de ADN. Estos fragmentos pueden ser ADN complementario u oligonucleótidos, y contienen secuencias que corresponden a fragmentos de ADN del organismo con el que se quiere trabajar. Estas secuencias son fijadas a la superficie sólida que contendrá el *microarray* en forma de puntos ordenados en forma de matriz bi o tridimensional. La superficie sobre la que se depositan suele ser de materiales semiconductores o de cristal. Posteriormente se aísla ARN o ADN de la muestra con la que se quiere trabajar. Estas secuencias se marcan posteriormente mediante fluorescencia. Una vez construido el *microarray* con las secuencias, y la muestra procesada y marcada, se hibridan ambos. Las secuencias del *microarray* y de la muestra que sean complementarias se unen, conteniendo los puntos donde se encuentran mayor proporción de marcadores fluorescentes. El *microarray*, una vez hibridado, se escanea utilizando técnicas de imagen, mediante láseres y microscopios confocales, para chequear la presencia de estos puntos fluorescentes, y poder así realizar una medida indirecta de la abundancia de las secuencias en la muestra.

Esta técnica puede ser utilizada con varios fines, pero en el caso de la transcriptómica, el uso mayoritario de los *microarrays* es el de realizar un análisis de expresión génica[319]. Los fragmentos que conforman la matriz de puntos del *microarray* son ADN complementarios u oligonucleótidos correspondientes a partes de los genes de los cuales se quiere medir la expresión. El diseño de estas secuencias es clave, ya que lo ideal es utilizar secuencias que no aparezcan repetidas en ningún otro gen del organismo a analizar, o por el contrario la medida de abundancia corresponderá a la suma de las concentraciones de los genes que la contengan. Una vez diseñado el *microarray*, en este caso se extrae ARN mensajero de la muestra con la que se va a trabajar, que normalmente se procesa para obtener su ADN complementario y marcarlo mediante fluorescencia, tal y como puede observarse en la Figura 1.2. Finalmente se analiza la intensidad de la fluorescencia para ver la expresión de los genes.

Comparar concentraciones de genes entre diferentes muestras puede ser bastante complicado de esta forma, ya que el uso de diferentes reactivos, concentraciones de muestras, marcadores fluorescentes, y demás parámetros que pueden variar dentro del experimento, pueden dar lugar a medidas muy diferentes de intensidad. De todas formas, con la mejora de esta tecnología a lo largo de los años, se han ido fabricando chips y analizadores que permiten una cada vez mejor

reproducibilidad. Estos incluyen secuencias *spike-ins*[183] cuyo grado de hibridación es conocido, para poder normalizar los niveles de intensidad de los diferentes puntos del *microarray*. De esta forma es posible comparar niveles de expresión de genes en diferentes condiciones si se utiliza un mismo tipo de chip.

Para intentar solventar este problema de las comparaciones se desarrollaron los *microarrays* de dos colores[259], en los cuales se utiliza ARN mensajero de dos muestras a comparar, que se marca en cada uno de los casos con un marcador fluorescente diferente. Posteriormente este ARN mensajero se mezcla y se hibrida con el *microarray* igual que en los otros casos, pero el análisis de imagen posterior cambia, ya que existe la posibilidad de que ARN mensajero marcado con los dos diferentes fluorocromos haya hibridado en un mismo punto del *microarray*. En este caso hay que realizar análisis en diferentes longitudes de onda para ver por separado la intensidad de cada uno de los canales. Esta intensidad no se utiliza para realizar cuantificaciones absolutas, ya que la intensidad en los puntos del *microarray* está fijado por las dos diferentes condiciones, que compiten entre ellas por hibridar en el chip. En este caso se utilizan las proporciones entre los valores de intensidad de las condiciones para determinar en cuál de ellas un determinado gen está más o menos expresado.

La tecnología de *microarrays* sigue presente hoy en día debido a su robustez y un coste relativamente bajo, pero presenta una serie de limitaciones[42]. Primero de todo, se asume que la intensidad de fluorescencia medida en un punto del *microarray* es proporcional a la cantidad de ARN o ADN complementario presente en la muestra. Esto no es siempre así, ya que a altas concentraciones la señal se satura, y a concentraciones muy bajas se favorece que no se produzca la unión. Por lo tanto, la señal sólo es linealmente proporcional en un rango de concentraciones de ARN en la muestra. También puede ser problemático el diseño de secuencias para detectar los genes en el *microarray*, ya que sobre todo en organismos más complejos como los mamíferos, puede ser muy complicado encontrar secuencias únicas de un solo gen, cuando existen otros con una homología de secuencia muy alta con éste. Esta situación se agrava en genes con diferentes variantes de *splicing*. Por último, el análisis de expresión en *microarrays* se limita a genes conocidos, ya que sólo se detectan genes de los cuales se conoce la secuencia y se puede generar un oligo o ADN complementario que pueda hibridar con el ARN de ese gen. Esto puede ser bastante problemático en genomas de especies que pueden ser altamente variables.

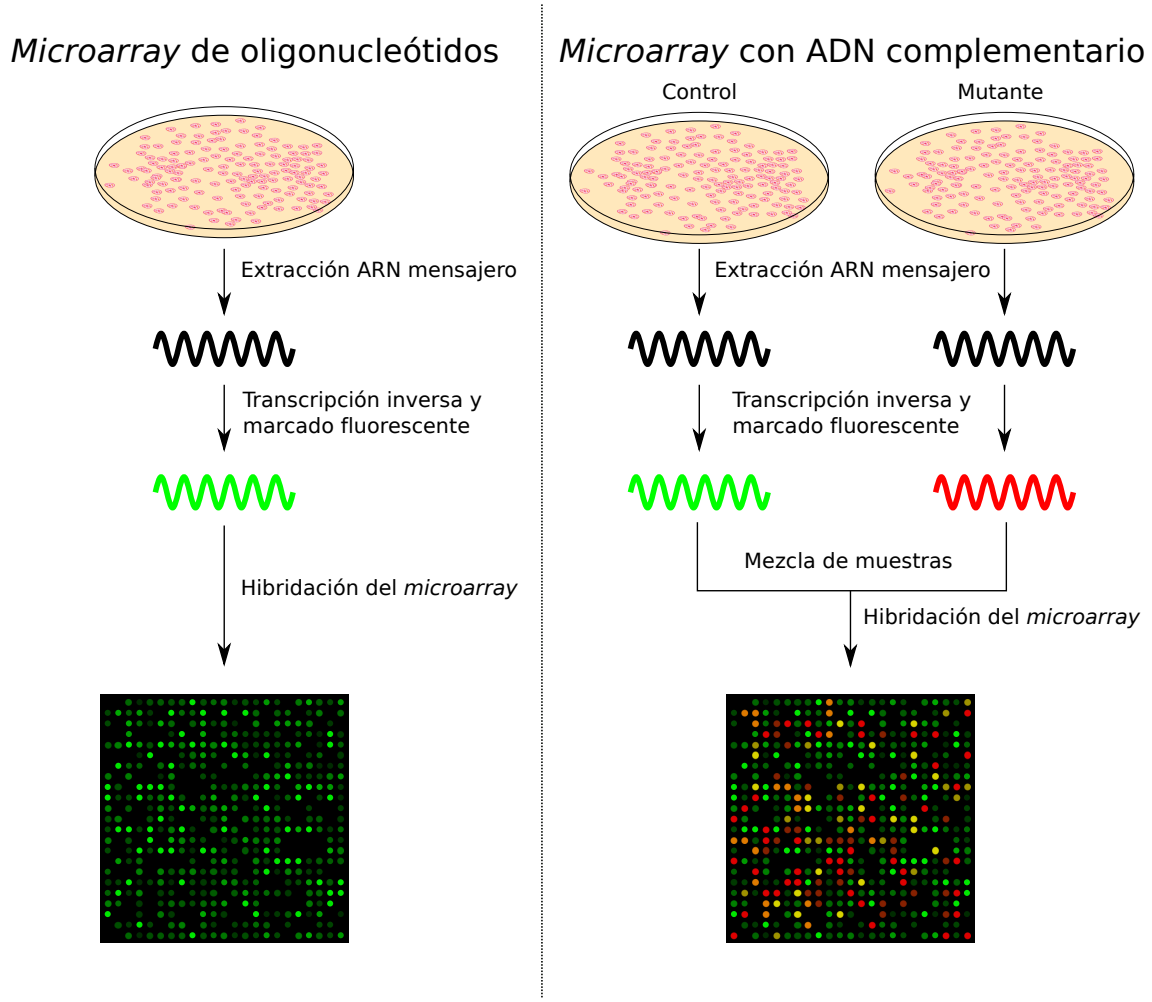


Figura 1.2: Esquema de análisis de diferentes tipos de *microarrays*.

1.1.2.1.2 *High-Throughput Sequencing*

Los métodos de determinación directa de secuencias, como el método de secuenciación de *Sanger*[249] existen desde mucho tiempo antes de la aparición de la ultra-secuenciación. Pero estos métodos eran muy lentos y caros, además de que generalmente no se podían utilizar para cuantificar ARN mensajero y medir expresión de genes. Por dar algunas cifras acerca del alto coste de estos métodos, la secuenciación del primer borrador del genoma humano en 2001[150], que fue completada mediante secuenciación de *Sanger*, tuvo un coste estimado de entre 500 y 1000 millones de dólares. Posteriormente se idearon técnicas de secuenciación de tránscritos basada en etiquetas como *CAGE*[140] y *SAGE*[294], que permitían realizar cuantificación de secuencias utilizando esta tecnología de secuenciación de *Sanger*, pero con las mismas limita-

ciones de coste y tiempo. Estas tecnologías sirvieron principalmente para anotar la estructura de los tránscritos más que para poder hacer cuantificaciones.

A finales de los 90 empezaron a desarrollarse diferentes técnicas de secuenciación de alto rendimiento[243] que empezaron a estar disponibles durante la primera década de los 2000, que se conocen como *High-Throughput Sequencing* o *Next Generation Sequencing*, y que revolucionaron el área de la biología. Estas técnicas mejoraron en varios órdenes de magnitud la cantidad de secuencias que se podían analizar en la misma cantidad de tiempo, además de que se redujeron considerablemente los costes derivados de la tarea. Esta tecnología es la predominante hoy en día en el mundo de la secuenciación, habiéndose mejorado considerablemente en la profundidad de la secuenciación, tiempo necesario y costes de la misma. Desde el inicio de esta tecnología, se buscó superar el límite económico de secuenciar un genoma humano por menos de 1000 dólares[176, 257], augurando una nueva era en la medicina personalizada. En el año 2014, la compañía *Illumina* lanzó al mercado el secuenciador *HiSeq X Ten*[108], que tiene la capacidad de secuenciar un genoma humano a una cobertura de *x30*, y que está cerca de bajar de esa barrera de los 1000 dólares.

Varias empresas desarrollaron métodos diferentes dentro del ámbito de la tecnología del *High-Throughput Sequencing*: *Illumina*, *Roche 454* e *Ion Torrent* entre otras. A pesar de las diferencias técnicas de cada una de estas aproximaciones, todas se basan en la capacidad de miniaturizar y paralelizar muchas operaciones de secuenciación, siendo capaces de leer millones de fragmentos de ADN/ARN en horas. Cada una de estas operaciones de secuenciación se alimentan con moléculas de ADN de tal forma que sólo haya una única molécula por operación, la cual se amplifica y entonces se secuencia. Estos aparatos permiten que cada una de estas operaciones se realice en un espacio muy pequeño, y permitiendo que se realicen grandes cantidades de estas operaciones en paralelo. Otra de las grandes ventajas de esta técnica es que se utilizan pequeños fragmentos de ADN sin ningún requerimiento adicional de inserción de plásmidos u otros vectores, permitiendo un ahorro adicional en tiempo. La secuenciación consta de varias fases, una primera de preparación de una librería de fragmentos de ADN, la fijación de los fragmentos, secuenciación y análisis de imagen.

La preparación de la librería es un paso fundamental, ya que tiene que producirse un extracto representativo y no sesgado del material genético que se va a analizar. Generalmente se utiliza una muestra de ADN que es fragmentado en trozos pequeños, y posteriormente estos fragmentos se inmovilizan en una superficie sólida o soporte. Esta inmovilización permite que

se produzcan millones de reacciones de secuenciación a la vez. En la mayor parte de las tecnologías, una vez fragmentado el ADN, este necesita ser amplificado para poder ser detectado de forma eficiente en el proceso de la secuenciación. Hay dos técnicas principalmente para realizar esto: la emulsión *PCR*[66] y la amplificación de fase sólida[77]. En la emulsión *PCR*, una vez fragmentado el ADN se liga con *primers PCR*. Después de hacer esto, se separan las dos hebras de ADN y éstas son capturadas por microesferas, bajo unas condiciones que propician que haya una molécula de ADN por microesfera. Posteriormente se amplifica la secuencia por *PCR*, manteniendose las copias de la misma pegadas a la microesfera. Por último, estas microesferas son o bien colocadas en una placa con celdillas, entrando cada microesfera en una de ellas, o fijadas químicamente a una superficie de cristal. La amplificación de fase sólida consiste en la ligación de unos adaptadores de los fragmentos, para que de esta forma se puedan pegar en una superficie. Posteriormente estas secuencias pegadas a la superficie se amplifican por *PCR*, generándose unas estructuras llamadas "*clusters* de ADN". Existen otras tecnologías, llamadas de *single-molecule sequencing*[104, 69], que permiten generar librerías sin amplificar las secuencias, requiriendo de menor cantidad de ADN. De esta forma se evitan mutaciones propiciadas por la amplificación. Hay varias aproximaciones, en las cuales pueden permanecer fijados a una superficie unos *primers*, las propias moléculas de ADN de la librería, o incluso unas polimerasas.

La forma de secuenciar en cada uno de estos casos puede ser muy diferente. En las muestras que se amplifican por *PCR*, el análisis de imagen se basa en medir el consenso de cada una de las secuencias amplificadas. Al irse añadiendo nucleótidos fluorescentes de uno en uno en cada paso de secuenciación, se necesita que este paso sea muy fiable, ya que de no serlo resulta en un desfase que provoca ruido en el análisis de fluorescencia. Al utilizar librerías *single-molecule sequencing* no existe este problema, pero podría darse el caso de que se añadieran múltiples bases en un ciclo, o que estos nucleótidos no tengan una etiqueta fluorescente.

El método de secuenciación de terminador cíclico[196] es un método cíclico que incorpora cada vez un nucleótido, se realiza un análisis de imagen por fluorescencia y una ruptura. En cada paso, se añaden bases nucleotídicas marcadas con fluorescencia, que además tienen un terminador reversible, de forma que cuando una de estas bases se une al ADN de los fragmentos, se pare la reacción. Una vez terminada esta fase, una cámara recoge la fluorescencia de estos nucleótidos para determinar cuál de ellos se ha unido. Posteriormente se realiza un lavado de las bases que no se han unido a ningún fragmento, y se elimina el terminador de las bases que se han unido. El proceso se repite múltiples veces leyendo un nucleótido de cada fragmento en

cada ciclo, obteniendo así las secuencias de la muestra.

El método de secuenciación por ligación utiliza una ligasa en vez de una polimerasa, y sondas marcadas con fluorescencia. En este caso, una sonda hibrida con su secuencia complementaria, que es un adaptador que se encuentra pegado a la cadena que se va a secuenciar. Entonces se añade una ADN ligasa y sondas marcadas con fluorescencia, uniéndose estos últimos a la primera sonda mediante la ligasa, reconociendo dos nucleótidos cada vez mediante fluorescencia después de realizar un lavado para eliminar las sondas no ligadas. Este ciclo se repite varias veces hasta completar la secuencia, y una vez hecho esto, se resetea la primera sonda, eliminándola y uniendo otra desplazada una base, y repitiendo todo el proceso varias veces para cubrir toda la secuencia.

El método de pirosecuenciación[242] está basado en la bioluminiscencia, ya que mide la liberación de pirofosfatos inorgánicos generándose luz visible mediante reacciones enzimáticas. En este caso no se utilizan nucleótidos modificados para parar la síntesis de ADN, sino que se manipula la ADN polimerasa añadiendo *dNTP*. Después de incorporarse el *dNTP*, la ADN polimerasa extiende la sonda y para. Se vuelve a añadir un *dNTP* complementario en el siguiente ciclo, reiniciándose la síntesis de ADN. La cantidad de luz emitida se graba en diagramas de flujo, que permiten reconstruir la secuencia de ADN.

Existe una tecnología de secuenciación basada en semiconductores[106] que es muy similar a la pirosecuenciación, pero en vez de hacer un análisis de luz emitida, como al añadirse un *dNTP* se libera un protón que hace decrecer el *pH*, se miden los cambios de *pH* para cada pocillo donde se encuentran las secuencias para determinar qué base se ha unido. Al igual que en la pirosecuenciación, tras medir el *pH*, los *dNTPs* se lavan, y el ciclo se repite.

Los métodos de secuenciación en tiempo real[69] permiten que no se interrumpa el proceso de síntesis de ADN. En este caso se van añadiendo continuamente *dNTPs* marcados con fluorescencia a la vez que se va realizando un análisis de imagen para detectar la fluorescencia del ADN sintetizado. Las moléculas de la ADN polimerasa se encuentran fijadas a la superficie de unos detectores llamados *Zero Mode Waveguide* (*ZMW*). Estos detectores evitan que la luz visible del láser lo atraviese completamente, iluminando un fondo de 30nm. Esto permite que los nucleótidos marcados por encima del detector no emitan luz, pudiéndose detectar la incorporación del nuevo nucleótido.

Los instrumentos de secuenciación comercializados hoy en día combinan alguno de estos métodos de preparación de librerías con alguna de las tecnologías de secuenciación explicadas anteriormente. La tecnología de *Illumina*[28], representada en la Figura 1.3, que permite obtener lecturas de fragmentos de unos 100-150 pares de bases, utiliza el proceso de amplificación de fase sólida y el método de terminación cíclica reversible utilizando cuatro colores. La tecnología de secuenciación de *Roche 454*[177], permite obtener lecturas mucho más largas que con *Illumina*. Utiliza el proceso de amplificación de gel *PCR* y el método de pirosecuenciación. *Ion Torrent*[228, 194] también utiliza el proceso de amplificación de gel *PCR* pero en este caso la tecnología de secuenciación basada en semiconductores. La plataforma *SOLiD*[289] utiliza la amplificación por gel *PCR* y el método de secuenciación por ligación. Los secuenciadores de *Pacific Biosciences*[69] utilizan la tecnología de *single-molecule sequencing* y la secuenciación en tiempo real, permitiendo adquirir lecturas de más de 10000 pares de bases. La plataforma de *Helicos BioSciences*[282] por su parte, también utiliza la tecnología de *single-molecule sequencing* pero acoplada al uso de la secuenciación por terminación cíclica reversible. De todos estos tipos de secuenciación, la más utilizada y que se ha impuesto finalmente ha sido la de *Illumina*[62], que proporciona una gran reproducibilidad y una buena calidad de lecturas, con una tasa de error cercana al 0.1 %[92].

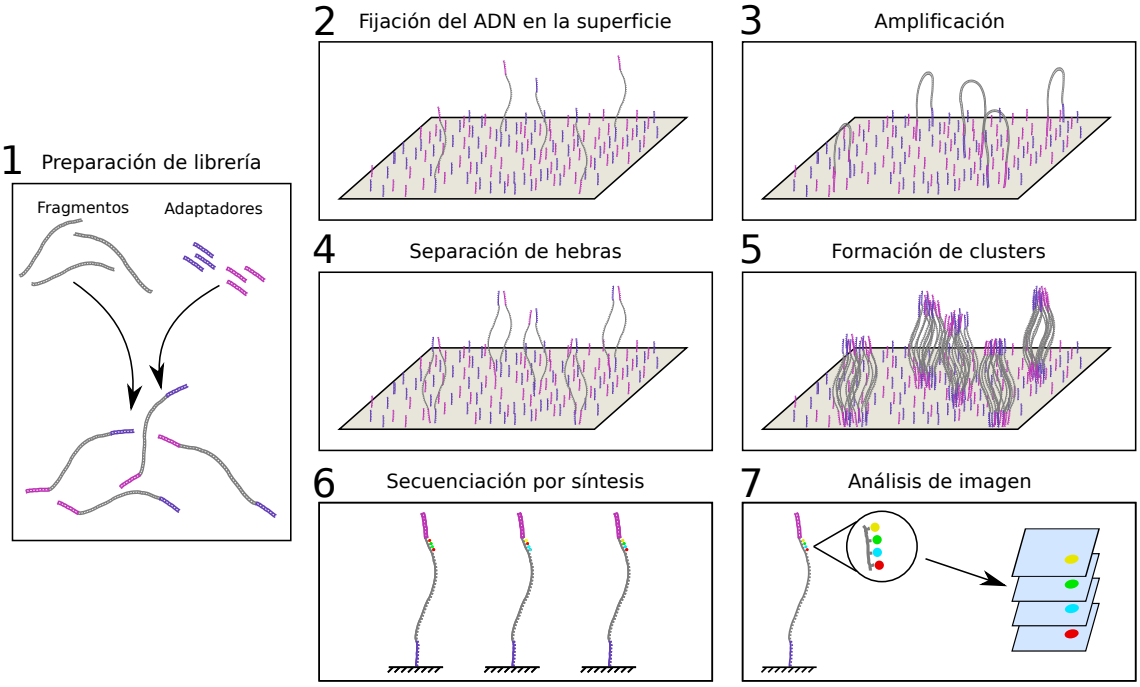


Figura 1.3: Descripción de la tecnología de secuenciación de *Illumina*.

A pesar de la novedad del *High-Throughput Sequencing*, esta tecnología ya ha sufrido algunas mejoras en los últimos años. Una de las más grandes fue el desarrollo de la secuenciación *paired-end*[84]. Con este método se secuencian los dos extremos de los fragmentos de ADN de la librería, pudiendo de esta manera producirse de una forma muy sencilla el doble de lecturas para la secuenciación, además de proporcionar una mejor precisión, ya que sabiendo el tamaño del fragmento, se puede inferir la posición relativa entre las dos lecturas de la pareja. Además de esta forma es mucho más fácil identificar duplicados por *PCR*, ya que es mucho más complicado encontrar parejas de lecturas idénticas en este tipo de secuenciación, asegurando casi por completo la existencia de duplicados *PCR* en el caso en el que aparezcan. La profundidad de secuenciación es modificable aumentando o disminuyendo el número de lecturas analizadas, pudiendo además restringirse la zona de secuenciación a una región de ADN muy pequeña[195], permitiendo así coberturas altísimas de la misma que permitan encontrar cambios mucho más sutiles entre diferentes muestras. En este sentido, al poder tener una cantidad muy alta de lecturas analizadas en una secuenciación, se puede utilizar para una única muestra, o aprovechar e introducir diferentes muestras en la librería mediante adaptadores de multiplexación[269]. Una vez obtenidas las lecturas, éstas se pueden filtrar fácilmente por estos adaptadores y separar las muestras. De esta forma se puede reducir el tiempo de secuenciación de estudios con multitud de muestras.

Gracias a esta tecnología se han podido conseguir grandes avances en el área de la transcriptómica, permitiendo comparar las abundancias relativas de los tránscritos de un organismo en diferentes condiciones, eliminando los problemas de saturación existentes en los *microarrays*, y no sólo limitándose a cuantificar los tránscritos de genes ya conocidos, sino permitiendo caracterizar nuevos tránscritos, isoformas e incluso transcriptomas completos de organismos cuyo genoma es conocido pero no así el transcriptoma.

1.1.2.2. Flujos de trabajo

A pesar de que los *microarrays* siguen utilizándose en la actualidad debido a su bajo coste, la investigación en transcriptómica se ha movido mayoritariamente a la utilización de *High-Throughput Sequencing*. El término usado para la utilización de esta tecnología para mapear y cuantificar transcriptomas se denomina *RNA-Seq*[302, 307]. Mediante estudios utilizando esta tecnología se puede medir la expresión génica de una muestra, pero además se pueden descubrir nuevos tránscritos y sitios de *splicing* alternativos[222, 275].

1.1.2.2.1 Preparación de muestras y secuenciación

Existen una gran variedad de aproximaciones en la preparación de muestras para su utilización en estudios de *RNA-Seq* debido a la cantidad de tecnologías diferentes, ya explicadas anteriormente. Pero hay partes clave del procedimiento que son comunes a todas las tecnologías.

La cantidad de ARN necesario para poder ser secuenciado varía de una plataforma a otra, pero existe la posibilidad de realizar una amplificación del mismo[290], existiendo la posibilidad de introducir ruido en la muestra. La mayor parte del ARN de la célula viene dado en forma de ARN ribosómico, el cual debe ser filtrado si queremos estudiar de forma completa la diversidad del transcriptoma presente en el resto del ARN. Una forma de hacer esto consiste en enriquecer de forma selectiva el ARN mensajero de la muestra, utilizando kits de eliminación de ARN ribosómico, o mediante el enriquecimiento de cadenas con colas poli(A), presentes sólo en los mensajeros. Los análisis de *RNA-Seq* enfocados en tipos específicos de ARN como ARN pequeño, ARN no codificante o micro ARN también necesitarían de un paso adicional de aislado de las moléculas. Una vez hecho esto, el ARN resultante tiene que ser ligado con *primers* para poder realizar un paso de transcripción inversa, y generar ADN complementario a estas cadenas de ARN, ya que las plataformas de secuenciación trabajan con ADN, no con ARN. Puede realizarse un paso adicional en el proceso de generación del ADN complementario para mantener la información sobre la hebra de la que procede el ARN[157], muy conveniente en el caso del estudio de transcriptomas complejos. El ADN resultante no puede ser introducido directamente en el secuenciador, ya que como se ha explicado previamente, este necesita ser preprocesado para generar una librería. Dependiendo de la tecnología usada, los pasos a seguir son diferentes, pero el resultado de la secuenciación es siempre el mismo, ficheros que contienen multitud de secuencias cortas con una calidad asociada a las mismas. El formato de ficheros *FASTQ*[52] nace para estandarizar la salida de la secuenciación con este tipo de tecnologías. Es un formato de texto *ASCII* basado en el formato *FASTA*[226] que permite almacenar las secuencias y sus calidades de una forma sencilla. Utiliza cuatro líneas por secuencia. La primera comienza siempre con el carácter @, seguido de un identificador de secuencia y una descripción opcional. La segunda línea contiene la secuencia en sí misma. La tercera línea empieza con un carácter + seguido opcionalmente del mismo identificador de secuencia y descripción. La cuarta línea contiene los valores de calidad de la secuencia de la segunda línea, por lo que debe contener el mismo número de caracteres. Los valores de calidad de la secuencia están codificados en la escala *Phred*[75, 76], que está relacionada de forma logarítmica con la probabilidad

de error en la asignación de nucleótidos:

$$Q_{PHRED} = -10 \log_{10}(P_e) \tag{1.1}$$

Esta escala, utilizando un único carácter por cada nucleótido, proporciona una forma simple y eficiente a nivel de espacio de almacenar la probabilidad de error de asignación. Existen dos bases *ASCII* principalmente utilizadas para codificar estos valores de calidad en los ficheros *FASTQ*, que suelen ir de un valor de Q de 0 hasta 40. Estas bases son 33 y 64, y representan el valor *ASCII* con un valor Q de 0, utilizando caracteres *ASCII* de la base añadiendo el valor Q para representar la calidad de cada nucleótido.

1.1.2.2.2 Procesamiento de secuencias

1.1.2.2.2.1. Control de calidad Una vez obtenido un fichero *FASTQ* con las secuencias provenientes de una muestra, es conveniente realizar un control previo sobre la calidad del mismo. Existe la posibilidad de que se hayan secuenciado partes de adaptadores, que exista algún tipo de contaminación en la muestra, o incluso que la secuenciación no haya sido de la calidad esperada por algún problema técnico. Existen programas tales como *FASTQC*[13], *NGSQC Toolkit*[55] o *PRINSEQ*[251] que permiten visualizar estadísticas de calidad de la secuenciación, comprobar la existencia de secuencias duplicadas con nucleótidos indeterminados (Ns), que implican problemas técnicos en la secuenciación, e incluso lanzar búsquedas de secuencias de adaptadores o de partes de genomas de organismos que podrían aparecer como contaminantes contra el fichero *FASTQ* para ver si existe algún tipo de problema con el mismo. Algunos de estos programas mencionados anteriormente permiten además filtrar las secuencias problemáticas o de baja calidad. Existen herramientas complementarias a los programas que no permiten realizar el filtrado, tales como *FASTX-Toolkit*[227, 147] o *TagCleaner*[252], capaces de procesar este tipo de ficheros. Especial cuidado merecen los *FASTQ* provenientes de una secuenciación *paired-end*, que normalmente aparecen como dos ficheros *FASTQ* que contienen cada uno un extremo del fragmento de cada secuencia, y que deben ser filtrados en paralelo, de tal forma que si se descarta uno de los extremos del fragmento, el otro también debe ser descartado, o separado en otro fichero. Esto es así porque muchos programas que permiten trabajar con secuencias *paired-end*[84] requieren que existan el mismo número de lecturas en cada uno de los dos ficheros, y en el mismo orden.

1.1.2.2.2.2. Alineamiento o ensamblaje *de novo* Una vez comprobada la calidad de los ficheros *FASTQ* y filtradas las posibles secuencias de mala calidad o contaminantes, el siguiente paso sería realizar el alineamiento contra un genoma de referencia o realizar un ensamblaje *de novo*[284, 44], tal y como puede apreciarse en la Figura 1.4. Elegir una u otra opción tiene unas implicaciones en coste y tiempo muy diferentes. En organismos cuyo genoma ya ha sido secuenciado, y de los cuales se tiene suficiente información acerca de sus genes, suele ser preferible realizar un alineamiento contra la referencia, que es bastante menos costoso computacionalmente hablando que un ensamblaje. Esta aproximación tiene algunas limitaciones, como no poder lidiar bien con lecturas de zonas repetitivas, o regiones que no existen en la referencia pero sí en la muestra. Por el contrario, los ensamblajes *de novo* permiten conocer mejor características estructurales de los tránscritos de una muestra en particular, pero requiriendo de unas necesidades computacionales muy grandes, no disponibles en muchos equipos de sobremesa actuales. Además de esto, los ensamblajes de novo suelen requerir de una profundidad de secuenciación muchísimo mayor para poder dar resultados de tránscritos consistentes. Este estudio se ha centrado en el análisis *RNA-Seq* mediante el alineamiento contra una referencia, por lo que conviene una explicación más exhaustiva de este tipo de análisis.

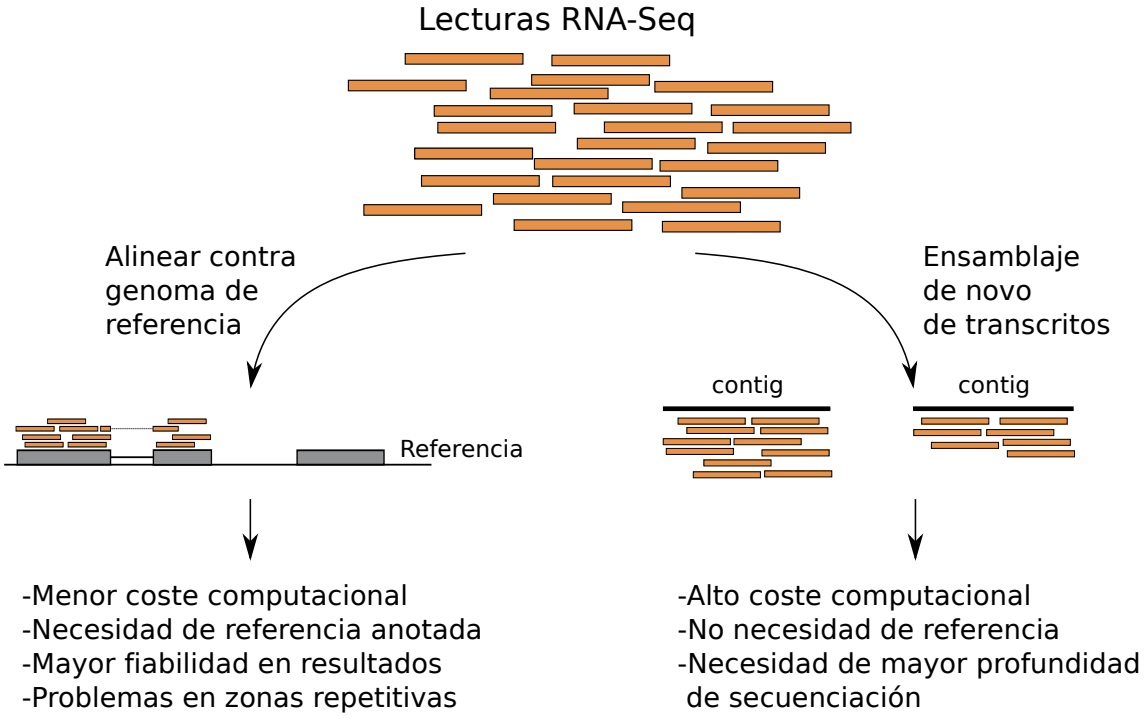


Figura 1.4: Diferentes formas de tratar lecturas de análisis *RNA-Seq* dependiendo de la existencia o no de una referencia.

Una de las mayores dificultades del alineamiento de lecturas de tránscritos contra una referencia consiste en mapear estas lecturas, que provienen básicamente de los exones de los genes, contra una secuencia genómica en la que también aparecen intrones. Existen programas de alineamiento de secuencias especificos para *RNA-Seq*, como *SOAPsplice*[123], *STAR*[63] o *TopHat*[284, 137], que son capaces de abordar específicamente este problema mediante una aproximación en dos fases, en la que primero se realiza un alineamiento de lecturas inicial para descubrir uniones de exones, que son utilizadas posteriormente para guiar un segundo alineamiento. Algunos de esos programas también son capaces de utilizar anotaciones de genes existentes para localizar estas uniones de exones y refinar aún más el alineamiento. También son capaces de priorizar los alineamientos en los cuales las lecturas de los dos extremos del fragmento en un análisis *paired-end* mapean de una forma consistente.

El resultado de un alineamiento suele ser un fichero en formato *SAM*[160] o *BAM*. En realidad son el mismo tipo de fichero, siendo el formato *SAM* un formato de texto tabulado, y *BAM* su versión binaria. Este tipo de ficheros consta de dos secciones, una cabecera que es opcional y cuyas líneas empiezan por el carácter @, y una sección de alineamiento. Cada línea de la cabecera contiene primero un código de dos letras seguido de información asociada al código, conteniendo datos referentes a la versión del formato, el orden y agrupamiento de los alineamientos, información sobre la referencia contra la que se ha alineado y sobre el *software* utilizado para el alineamiento. Cada línea de la sección de alineamiento contiene once campos con información sobre el propio alineamiento, tal como la posición exacta del mapeo, datos adicionales de calidad del alineamiento, las diferencias con la referencia, etc, y además puede contener un número variable de campos adicionales optativos, que suelen contener información específica del alineador. Existen muchos programas que requieren un cierto orden en este tipo de ficheros, por lo que existen herramientas que permiten manipularlos tales como *SAMtools*[160] o *Picard*[39]. Estas herramientas, además de modificar el orden de estos ficheros, permiten buscar y eliminar secuencias duplicadas, unir varios ficheros, generar índices para los mismos, convertir de formato de texto a binario y viceversa, y otras operaciones similares.

1.1.2.2.3 Análisis de secuencias

Una vez alineadas las secuencias, existen una serie de análisis que se pueden llevar a cabo con las mismas. Hay flujos de trabajo que son exclusivos de *RNA-Seq*, y hay otros que pertenecen a análisis de *High-Throughput Sequencing* de ADN y pueden adaptarse para trabajar con

ARN. A continuación se explican los tipos de análisis de interés para la presente investigación.

1.1.2.2.3.1. Análisis de expresión génica Para medir la expresión de los genes en un análisis *RNA-Seq* primero necesitamos conocer las lecturas que caen dentro de los límites de las zonas codificantes de cada gen o exones. Para esto existen paquetes de herramientas como *HTSeq*[12] o *featureCounts*[163] utilizando además ficheros *GFF* o *GTF* que contienen información sobre las coordenadas de genes y demás características de los mismos (comienzo y final de exones, intrones, *UTRs*...).

Existen diversas fuentes de variabilidad que necesitan de una normalización posterior. La fragmentación del ARN llevada a cabo en la construcción de la librería causa que tránscritos más largos generen más lecturas que los cortos teniendo ambos la misma abundancia[217]. Por otro lado, la diferencia en el número de lecturas generadas en cada secuenciación produce fluctuaciones en el número de lecturas mapeadas en diferentes muestras[178]. Por estos motivos aparecieron métodos de normalización, como el cálculo de *RPKM*[201] y *FPKM*, que normalizan el conteo de tránscritos teniendo en cuenta la longitud de los mismos y la abundancia total de lecturas de la muestra:

$$R = \frac{10^9 C}{NL} \tag{1.2}$$

Siendo *C* el número de lecturas que caen en los exones de determinado gen, *N* el número total de lecturas que han caído en cualquier exón dentro del experimento, y *L* la longitud de la suma de los exones en pares de bases. La única diferencia entre *RPKM* y *FPKM* es que en el primero se cuentan lecturas, mientras que en el segundo secuentan fragmentos. El número de lecturas y fragmentos es diferente únicamente en estudios en los que se utilizan lecturas *paired-end*. A pesar de ser ampliamente utilizados, estos métodos han sido muy criticados, y se han propuesto nuevos métodos como el cálculo de *TPM*[299], en el cual primero se normaliza por la longitud del gen, y posteriormente se normaliza por la profundidad de la secuenciación. A pesar de todos estos métodos de normalización, es muy complicado en casos de genes con conteos de lecturas muy bajos saber si realmente estas lecturas se deben a expresión real del gen o son artefactos debido a algún tipo de contaminación de la muestra o ruido.

Cuando ya se ha estimado un valor de expresión normalizado para cada gen, una pregunta importante consiste en entender cómo estos niveles de expresión cambian en diferentes condiciones. Con la aparición de los *microarrays* comenzó el desarrollo de metodologías para el análisis estadístico de su expresión diferencial[287, 49, 265]. Estas metodologías son aplica-

bles directamente a los datos de expresión de *RNA-Seq*, cambiando los valores de intensidad de fluorescencia por los valores normalizados de abundancia de tránscritos provenientes del conteo de lecturas. En un comienzo, los métodos intentaban modelar estas lecturas utilizando distribuciones como la de *Poisson*[178], pero se ha visto que estas distribuciones no tienen en cuenta la variabilidad biológica entre muestras[241]. Por esta razón, y teniendo en cuenta que muchas veces no existe un número suficiente de réplicas biológicas para inferir esta variabilidad, muchos métodos tales como *DESeq*[169], *EdgeR*[185] o *Cuffdiff*[283] intentan modelar esta variabilidad en el conteo utilizando aproximaciones paramétricas como por ejemplo la distribución binomial negativa. Estos métodos cuentan con paquetes de herramientas que normalmente aceptan directamente tablas de entrada con las lecturas de los genes en crudo, ya que cada uno de ellos normaliza estas lecturas con un método propio. Una vez realizado el análisis, los resultados suelen consistir en tablas con genes, y un valor de *fold change*, p-valor y q-valor referente a un *FDR* asociado a cada uno de ellos. Filtrando estos datos se pueden obtener listas de genes que han variado significativamente entre diferentes condiciones de unas muestras.

1.1.2.2.3.2. Análisis de variantes La búsqueda de mutaciones en una muestra determinada y la búsqueda de significación de las mismas a efectos fenotípicos es otro de los análisis más comunes trabajando con datos de *High-Throughput Sequencing*. De hecho, el comienzo de su utilización auguró un futuro en el que la medicina personalizada[127], a través de la toma de una muestra del paciente, y su análisis mediante secuenciación, permitiera conocer el origen genético de muchas de nuestras enfermedades, y poder actuar en consecuencia. Los avances de los últimos años han permitido que esto empiece a ser una realidad[253], aunque la tecnología es todavía limitada y cara. El estudio de poblaciones también se ha visto beneficiado por este tipo de análisis, permitiendo de una forma mucho más sencilla la realización de ensayos de asociación de genomas completos[296].

Debido a que en este caso se trata de un análisis de mutaciones en el ADN, el análisis de variantes en un estudio de *RNA-Seq* queda limitado a las zonas codificantes del genoma que además se han expresado en una determinada muestra. Por lo tanto, puede ser útil en algunos casos, pero para poder obtener una lista más completa de mutaciones de una muestra es necesario secuenciar ADN nuclear.

Para este tipo de análisis se parte del alineamiento de las secuencias de una determinada muestra. A la hora de realizar los alineamientos hay que tener en cuenta que la muestra puede

contener una serie de mutaciones, y es necesario incluir esta posibilidad como parámetro, en otro caso sólo tendríamos alineamientos perfectos contra la secuencia del genoma de referencia y estaríamos perdiendo la información acerca de las variantes. Son las lecturas con alineamientos imperfectos contra la referencia las que son útiles a la hora de descubrir mutaciones en la muestra. Existen programas, como *Samtools*[160] con su herramienta *mpileup*, o *GATK*[186] que, a partir del alineamiento, buscan posiciones donde exista una cantidad suficiente de lecturas en las cuales, en un mismo nucleótido, exista una divergencia con respecto a la referencia, como puede observarse en la Figura 1.5. A cada posible mutación se le asigna una puntuación dependiendo del número de lecturas que soportan esa evidencia, y de la calidad del alineamiento de esas lecturas contra la referencia, entre otros factores. Las posibles mutaciones que tienen una calidad baja se filtran, y las que se mantienen se escriben en un fichero de texto denominado *VCF*[292]. Este tipo de ficheros contienen tres partes: una de meta-información, en la que cada una de las líneas comienza con *##*, y que contiene básicamente datos sobre la versión del formato de fichero, el programa utilizado para generarlo, la referencia contra la que se ha alineado el genoma, e información sobre el formato usado en el genotipado; una línea de cabecera, que comienza con *#*; y líneas con datos, que contienen información sobre las posiciones donde existe la variación, e información de genotipado de las mismas, pudiendo aparecer varias muestras separadas por tabuladores. Existen programas tales como *snpEff*[50] o *VEP*[187] que permiten procesar estos ficheros *VCF* generados, pudiendo asociar las mutaciones por coordenadas a un determinado gen, o intentar ver el posible efecto de esa mutación, si está en una zona codificante, sobre la proteína que se generaría.

1.1.3. Proteómica

Una vez que el ADN se ha transcrito en ARN mensajero en el núcleo de la célula, éste sale al citoplasma y llega a los ribosomas, que son complejos macromoleculares de proteínas y ARN encargados de sintetizar proteínas a partir de la información contenida en el mensajero. Estas proteínas, que son moléculas formadas por cadenas de aminoácidos, desempeñan un papel fundamental para la vida, ya que son responsables de gran parte de los procesos esenciales que se producen en un organismo, tales como el transporte de moléculas, interviniendo en reacciones bioquímicas o tienen función estructural. El conjunto de proteínas expresadas en una condición determinada es el proteoma, y la proteómica es el área encargada de estudiarlo. El proteoma es un conjunto más variable que el transcriptoma, ya que depende por un lado de las características del mismo, que incluyen su secuencia, estructura tridimensional e interacciones entre

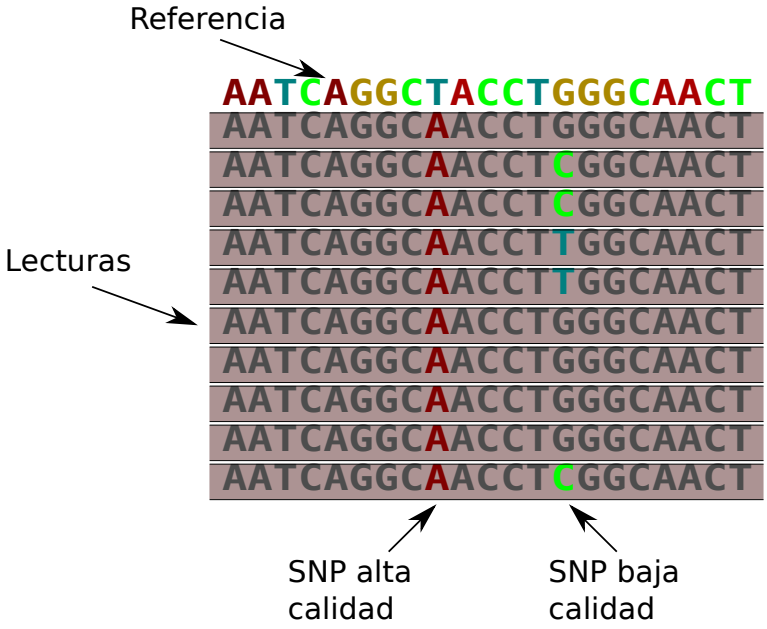


Figura 1.5: Apilamiento de lecturas en análisis de *SNPs* utilizando *NGS*.

diferentes proteínas o con otras moléculas, entre otros, y de otros factores como los cambios químicos ocurridos después de la síntesis de las proteínas llamados modificaciones postraduccionales (*PTMs*)[232]. El concepto de proteoma apareció por primera vez en el año 1994[303]. El desarrollo de este área ha venido en gran medida marcado por la aplicación en este ámbito de la tecnología de la espectrometría de masas[3]. Pero también existen otras tecnologías y recursos necesarios para llevar a cabo experimentos. Estas incluyen tecnologías de separación de proteínas, muy utilizadas como paso previo a la espectrometría de masas.

Estos avances tecnológicos han permitido el auge de estudios de descubrimiento de proteínas, muy importantes en el contexto del hallazgo de nuevos biomarcadores y en proyectos de ciencia básica. De esta forma han surgido proyectos como el *Human Proteome Project* (*HPP*)[155] perteneciente al *Human Proteome Organization* (*HUPO*), que consiste en un esfuerzo global colaborativo consistente en mapear y caracterizar todas las proteínas codificadas por genes humanos. La creación del mapa de la arquitectura molecular basada en proteínas del cuerpo humano puede ser muy útil para indagar sobre funciones biológicas y moleculares de las mismas y para mejorar en el diagnóstico y tratamiento de enfermedades. Aparte del trabajo colaborativo del *HPP*, dos grupos han avanzado notablemente en el estudio del proteoma, habiendo publicado *drafts* del proteoma humano[138, 308].

En este tipo de proyectos de descubrimiento, existe un gran problema, ya que debido a la

gran cantidad de tipos de células diferentes en un organismo, el proteoma expresado en un momento determinado puede ser muy diferente, siendo complicado conocer la forma de encontrar qué tipo de célula, y en qué condiciones determinadas puede expresar determinada proteína. De esta forma la bioinformática aparece, integrando datos de experimentos de diversa índole y construyendo bases de datos que sirvan como herramienta para poder realizar hipótesis en este sentido[288].

1.1.3.1. Identificación de proteínas mediante espectrometría de masas

1.1.3.1.1 Espectrómetros de masas

Un espectrómetro de masas[114] es un aparato capaz de medir con una gran precisión la masa de moléculas ionizadas, o más correctamente la relación masa-carga (*m/z*), en este caso de péptidos y proteínas. El proceso realizado por el espectrómetro de masas consta de varias partes que pueden variar, esquematizados en la Figura 1.6. A continuación se detallan los pasos más importantes.

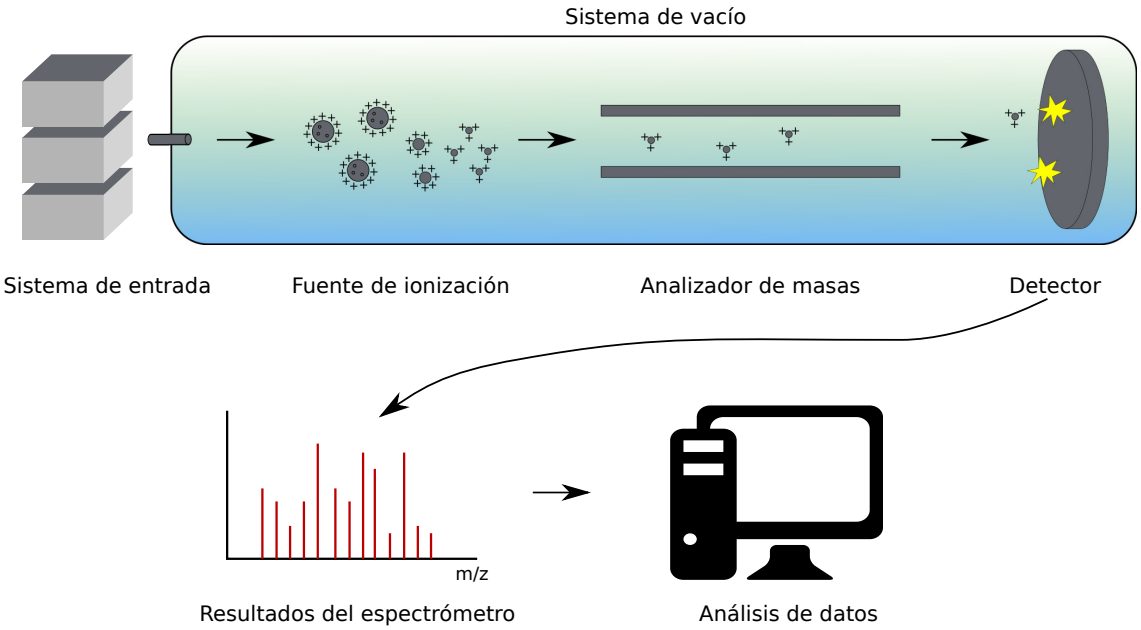


Figura 1.6: Esquema de las partes de un espectrómetro de masas y análisis posterior.

1.1.3.1.1.1. Entrada de la muestra Trabajando con proteínas, existe un primer nivel de separación previo, que puede ocurrir con un fraccionamiento subcelular o con separación de proteínas utilizando electroforesis en gel. Una vez la muestra se ha separado de esta forma, y se ha realizado una digestión utilizando enzimas proteasas con un patrón de corte conocido, como la tripsina, el siguiente paso de separación ocurre en un sistema de cromatografía de alta resolución como los sistemas de cromatografía líquida de alto rendimiento (*HPLC*), que suele estar acoplado a la entrada del espectrómetro de masas y permite que se introduzca gradualmente una muestra al espectrómetro. La *HPLC* es un tipo de cromatografía de columna en la cual se bombea a alta presión una muestra en un solvente (fase móvil) a través de una columna cromatográfica (fase estacionaria). La muestra se va introduciendo en pequeñas cantidades y sus componentes se retrasan diferencialmente en función de sus interacciones con la fase estacionaria a medida que atraviesan la columna. El tiempo que tarda la parte separada de la muestra en ser eluida de la columna se denomina tiempo de retención, el cual se utiliza como referencia para identificar el compuesto.

1.1.3.1.1.2. Ionización La muestra primero debe ser ionizada, para así volatilizar la muestra y poder introducirla en el sistema de vacío del espectrómetro. La ionización de péptidos y proteínas fue un problema en el comienzo del desarrollo de estas metodologías debido a que las técnicas existentes destruían las moléculas. Con el desarrollo de técnicas de ionización suave se permitió su ionización sin producir una fragmentación excesiva. Las técnicas *MALDI*[132] (*matrix-assisted laser desorption/ionization*) y *ESI*[78] (*electrospray ionization*) fueron claves para esto. La técnica *MALDI* consiste en primero mezclar la muestra con una matriz sólida que suele consistir en un material orgánico. A continuación la muestra se irradia con un láser pulsado, expulsándose entonces iones cargados de la matriz, cationes y parte de la muestra, generándose una nube gaseosa. Finalmente la muestra es ionizada por colisiones con las otras moléculas del gas y se acelera hacia la cámara de vacío del espectrómetro de masas. En la técnica *ESI* se introduce la muestra disuelta en un solvente por un capilar de metal cargado. Debido a la carga eléctrica, el líquido sale de la punta del capilar en forma de gotas, formándose un aerosol. El solvente se evapora, quedando los iones de la muestra, que entonces se dirigen hacia la cámara de vacío del espectrómetro.

1.1.3.1.1.3. Analizador de masas Una vez que la muestra se ha ionizado, los iones pasan al lugar donde se encuentra el analizador de masas. Éste se utiliza para separar los iones obtenidos de la muestra por su relación de masa y carga (*m/z*) y suele consistir en una cámara de vacío

donde se aplica algún tipo de campo eléctrico o magnético para poder ver las diferencias de comportamiento de los iones al moverse a través de la cámara. Existen varios tipos de analizadores de masas, pero los más utilizados en proteómica son los siguientes: analizadores *Time Of Flight* (*TOF*), que utilizan un campo eléctrico para acelerar los iones y midiendo el tiempo que toman para alcanzar el detector, se puede inferir la relación *m/z*; cuadrupolos, que constan de cuatro varillas de metal enfrentadas en pares, sobre los que se aplica una corriente continua y otra alterna, permitiendo esto crear un campo eléctrico controlado que desvía selectivamente los iones que lo atraviesan, pudiéndose filtrar de esta manera los iones con un margen muy pequeño de *m/z*, siendo éstos los únicos que llegarán al detector; y las trampas iónicas, que funcionan de un modo similar a los cuadrupolos, pero en vez de desviar los iones que no se encuentran en un determinado margen de *m/z*, éstos se confinan y almacenan en una cámara, y posteriormente son liberados de forma selectiva. Existen varios tipos de trampas iónicas, entre las cuales destacan las *Quadrupole Ion Trap*[254] (*QIT*) y *Orbitrap*[89].

1.1.3.1.1.4. Detector Acoplado con el analizador de masas es necesaria la existencia de un detector, que registra la carga inducida o la corriente producida cuando el ion pasa cerca o golpea una superficie. Los más utilizados se denominan multiplicadores de electrones, los cuales se componen de varias placas sobre las cuales se aplica una diferencia de potencial, que entonces producen una descarga de electrones cuando un ion incide sobre su superficie, produciéndose entonces un efecto cascada por la existencia de numerosas placas.

1.1.3.1.2 Aproximaciones de identificación de proteínas

Existen dos aproximaciones principales para la identificación de proteínas, la *top-down*[133], en la que se analizan proteínas intactas que suelen estar aisladas, y la *bottom-up*[318], en la cual se utilizan proteínas que son previamente fragmentadas mediante enzimas proteasas con un patrón de corte conocido, como la tripsina. En ambos casos, utilizando un espectrómetro de masas, es posible obtener un espectro de masa, que no es más que la huella de *m/z* para los cuales hay iones presentes en la muestra. El objetivo en proteómica consiste en interpretar estos espectros, asignando secuencias de péptidos a cada uno de ellos. En muestras de proteínas suficientemente aisladas con este paso es suficiente, pero en estudios de proteómica en los que se utilizan mezclas complejas, llamados normalmente como proteómica de *shotgun*, es necesario un paso adicional, ya que de no llevarse a cabo aparecen muchos espectros similares, los

cuales no pueden ser diferenciados. En el caso de la aproximación *top-down* puede no ser necesario este paso adicional debido al grado de aislamiento de la muestra, pero en aproximaciones *bottom-up*, que son las más utilizadas en proteómica de *shotgun*, es imprescindible para una identificación fiable.

1.1.3.1.3 Espectrometría de masas en tándem

La realización de este paso adicional en mezclas complejas se realiza mediante la espectrometría de masas en tándem[262] (*MS/MS* o *Tandem Mass Spectrometry*). Esta técnica incluye varios pasos de selección mediante espectrometría de masas con pasos intermedios de fragmentación. En el caso de proteínas consiste en que una vez los péptidos ya ionizados se encuentran dentro del analizador, se vuelven a fragmentar generando iones más pequeños que también son detectados, haciendo que el patrón de fragmentación sea mucho más específico de secuencia. Normalmente el espectrómetro primero registra los espectros de los péptidos ionizados, y selecciona los de mayor intensidad, llamados iones precursores, para después fragmentarlos y generar el espectro de fragmentación.

1.1.3.1.4 Análisis de datos

Una vez llevada a cabo la espectrometría de masas, para poder realizar el análisis de identificación de proteínas hay que procesar toda la información generada en forma de espectros. Dependiendo del espectrómetro, el formato de salida de estos ficheros de espectros puede ser muy diferente. Existen iniciativas, como la del *HUPO-PSI* que intentan estandarizar estos formatos, habiéndose creado de esta forma el estándar *MZml*[179], que contiene en un formato *XML* los diferentes datos extraídos del espectrómetro.

1.1.3.1.4.1. Motores de búsqueda de proteínas Los resultados de estos experimentos contienen espectros pertenecientes a mezclas de proteínas, siendo en proteómica de *shotgun*, que es de las más utilizadas en experimentos de alto rendimiento, especialmente complejas. De esta forma, se requiere de un paso de asignación de espectros a péptidos. Esta asignación se lleva a cabo mediante el uso de programas de búsqueda en bases de datos de secuencias[73] tal y como

se muestra en la Figura 1.7, en los cuales se introduce la lista de espectros obtenidos en el experimento, y una base de datos de secuencias de proteínas que normalmente corresponden con el conjunto de proteínas del organismo analizado. Estos programas realizan una digestión sintética de las proteínas, eligiendo el mismo tipo de enzima que se ha empleado en el experimento, para generar péptidos del mismo tipo. A partir de estos péptidos sintéticos se realiza un cálculo para determinar la *m/z* de cada uno de ellos, y entonces se compara el espectro del experimento con el de las *m/z* de estos péptidos sintéticos, asignándose puntuaciones a esas comparaciones en base a su grado de similitud. Para comprobar la fiabilidad de las asignaciones se realiza un test estadístico midiendo entre otros la tasa de falsos descubrimientos (*FDR*). Para ello se suele utilizar una base de datos señuelo (*decoy*)[72], donde los espectros del experimento son comparados con espectros teóricos provenientes de secuencias que no corresponden con ninguna proteína real. De esta forma el *FDR* se define como la proporción de asignaciones incorrectas (asignaciones a secuencias señuelo) que se aceptan utilizando un umbral de puntuación determinado. Hay varios métodos para obtener secuencias *decoy*[72, 209]. Entre los más populares destacan la generación de las secuencias de forma aleatoria, la inversión de las proteínas, o la realización de una digestión *in silico* de las proteínas seguida de una inversión de los péptidos obtenidos conservando el extremo C-terminal (pseudoinversa). Ejemplos de programas de bases de datos de búsqueda son *OMSSA*[88], *Mascot*[229], *X!tandem*[54] o *SEQUEST*[73]. Todos estos métodos son sumamente costosos computacionalmente, requiriendo de horas, o incluso días, para completar la tarea de asignación de espectros a péptidos.

1.1.3.1.4.2. Inferencia de proteínas Desde el momento en el que se fragmentan las proteínas con enzimas, todo el análisis se realiza a nivel de péptido. No es fácil, a partir de los péptidos identificados, inferir qué proteínas son las que aparecen en el análisis[210] debido a la redundancia de secuencias. Esta aparece por diferentes causas. Una de ellas es la existencia del *splicing* alternativo, que impide muchas veces la discriminación de la isoforma expresada debido a la posibilidad de que ninguno de los péptidos identificados sea único de alguna ellas[33]. Otro problema es la existencia de las familias de proteínas, cuyos miembros tienen un alto grado de homología entre sí, lo cual provoca que haya péptidos compartidos por diferentes miembros de la misma familia, lo que también dificulta la tarea de inferencia[235].

En muchos estudios no se describe la forma por la cual una proteína aparece como identificada. En ocasiones hay péptidos que aparecen como identificados en más de una entrada de la base de datos y se asignan de forma aleatoria a cualquiera de las proteínas de la identificación.

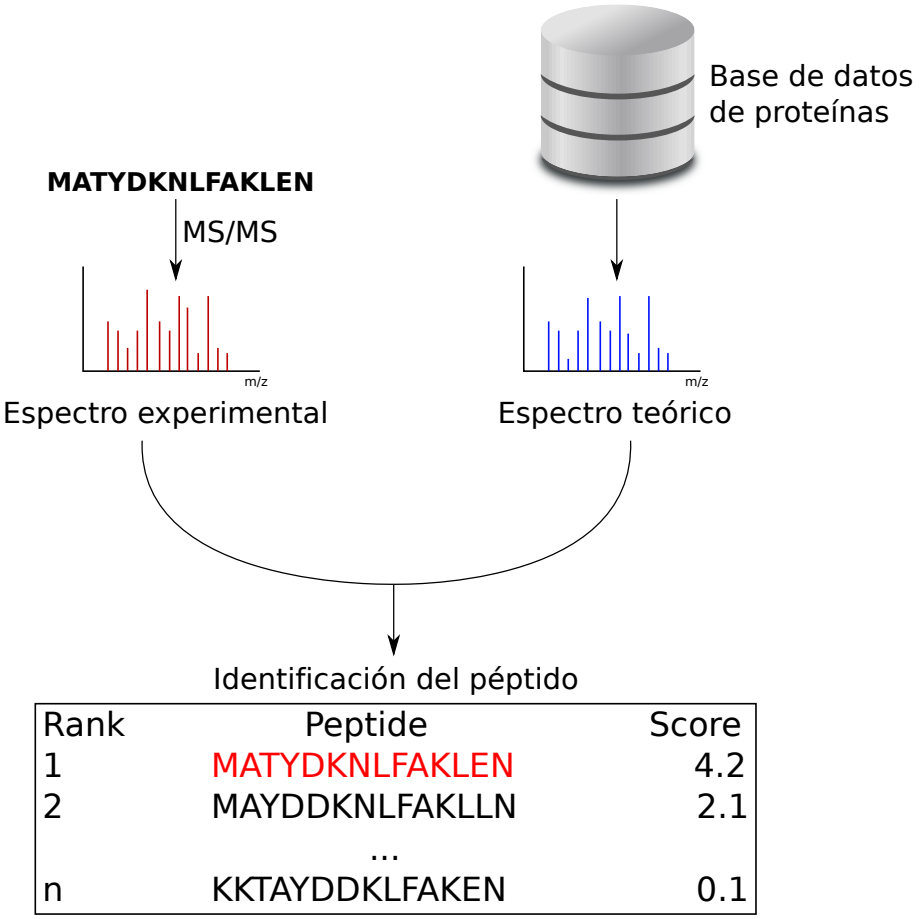


Figura 1.7: Esquema de funcionamiento de motor de búsqueda en proteómica de *shotgun*.

En otras ocasiones se toman como inferidas todas las proteínas de la identificación. De esta forma no sólo se sobreestima el número de proteínas identificadas, sino que además se puede realizar una interpretación biológica incorrecta. Por lo tanto, es necesario utilizar una metodología más precisa a la hora de realizar esta inferencia para poder interpretar los datos de una forma más correcta, como la explicada en [210].

Para poder clasificar los diferentes grados de inferencia de las proteínas, primero hay que clasificar los péptidos. Estos pueden ser únicos, los cuales son exclusivos de una única proteína; discriminantes, que son péptidos compartidos cuya presencia puede ser explicada por un grupo de proteínas que no contienen péptidos únicos; y no discriminantes, que son péptidos compartidos cuya presencia puede ser explicada por proteínas con péptidos únicos o discriminantes, básicamente péptidos que aparecen en diferentes grupos de proteínas. A partir de estos tipos de péptidos, las proteínas inferidas pueden clasificarse en cuatro grupos de evidencia: inferencia

de proteínas concluyentes, que contienen péptidos únicos; inferencia de proteínas no conclusivas, que sólo contienen péptidos no discriminantes; inferencia de proteínas indistinguibles, que son grupos de proteínas que comparten todos los péptidos, incluyendo alguno discriminante; y grupo ambiguo, que es un grupo de proteínas que explican la presencia de un grupo de péptidos discriminantes. Un ejemplo de esta inferencia de proteínas puede verse en la Figura 1.8. A partir de esta clasificación, y utilizando el principio de la navaja de *Occam*[211], se puede obtener una lista mínima de proteínas que explique la lista de péptidos identificados. Esta lista contendría todas las identificaciones de proteínas concluyentes, y agruparía las identificaciones de proteínas indistinguibles entre sí. Existen también métodos más refinados, como el propuesto en [233] que mejora la inferencia en caso de ambigüedad.

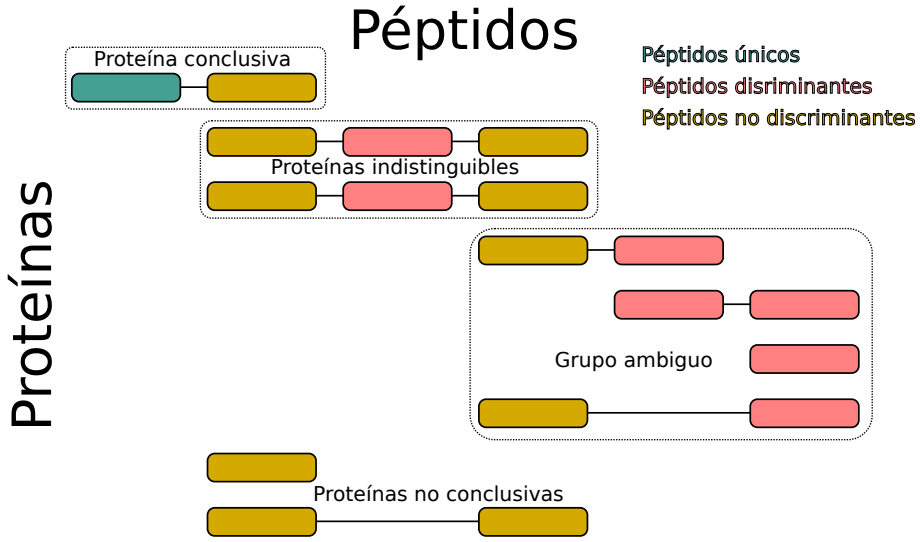


Figura 1.8: Ejemplo de inferencia de proteínas a partir de la clasificación previa de los péptidos.

1.1.4. Proteogenómica

En aproximaciones proteómicas *bottom-up* utilizando espectrometría de masas en tándem, las más utilizadas para estudios con mezclas complejas de proteínas, el resultado del análisis son espectros pertenecientes a péptidos y fragmentos de los mismos. Como se ha comentado anteriormente, estos datos se analizan generando espectros teóricos a partir de bases de datos de proteínas, y se comparan con los espectros reales del experimento. Pero en este tipo de estudios se asume que todas las secuencias que codifican proteínas en el genoma son conocidas y están correctamente anotadas, y que además existe una base de datos de secuencias de proteínas de

referencia para el mismo. Por lo tanto, la colección de proteínas de un organismo se considera un conjunto fijo. La proteómica hasta ahora ha permitido el estudio de la variabilidad de expresión de las proteínas e identificar modificaciones postraduccionales, entre otros.

El problema de hacer estas asunciones está en que muchos péptidos de las muestras no van a estar nunca presentes en estas bases de datos de proteínas basadas en la referencia, y por lo tanto no serían identificados en los flujos de trabajo de proteómica de *shotgun*. Esto puede ser debido a que los péptidos contengan mutaciones puntuales, provengan de proteínas con formas de *splicing* no detectadas, o incluso que provengan de proteínas no conocidas. Existe la posibilidad de realizar secuenciaciones *de novo* de los péptidos[170], pero la tasa de errores es muy alta, además de ser un proceso muy costoso computacionalmente.

Debido a estos problemas, y gracias a los avances en el campo de la transcriptómica, sobre todo gracias a la tecnología de *RNA-Seq*, aparece la proteogenómica[126]. En un inicio el término proteogenómica se utilizó en estudios de proteómica que fueron utilizados para mejorar la anotación del genoma. En definitiva, la proteogenómica es un campo en el que se utilizan datos de experimentos de genómica y proteómica, y se combinan para mejorar los resultados obtenidos en ambas. Actualmente se utiliza en gran medida, aprovechando los avances de la genómica y la transcriptómica, para estudiar las particularidades de la secuencia de un organismo, tal y como se muestra en la Figura 1.9, y poder de esta manera mejorar el proceso de asignación de espectros a péptidos. De esta forma, se pueden generar bases de datos de proteínas más completos con información específica de una muestra, y minimizar el número de espectros que quedan sin asignar.

1.1.4.1. Análisis de datos

En el contexto de la proteogenómica primero hay que realizar un experimento de *RNA-Seq* y otro de proteómica de *shotgun* sobre una misma muestra. La mayor parte del procesamiento de los datos que difiere de un análisis estándar consiste en la preparación de bases de datos a partir de los datos transcriptómicos. Los siguientes apartados describen este flujo de análisis.

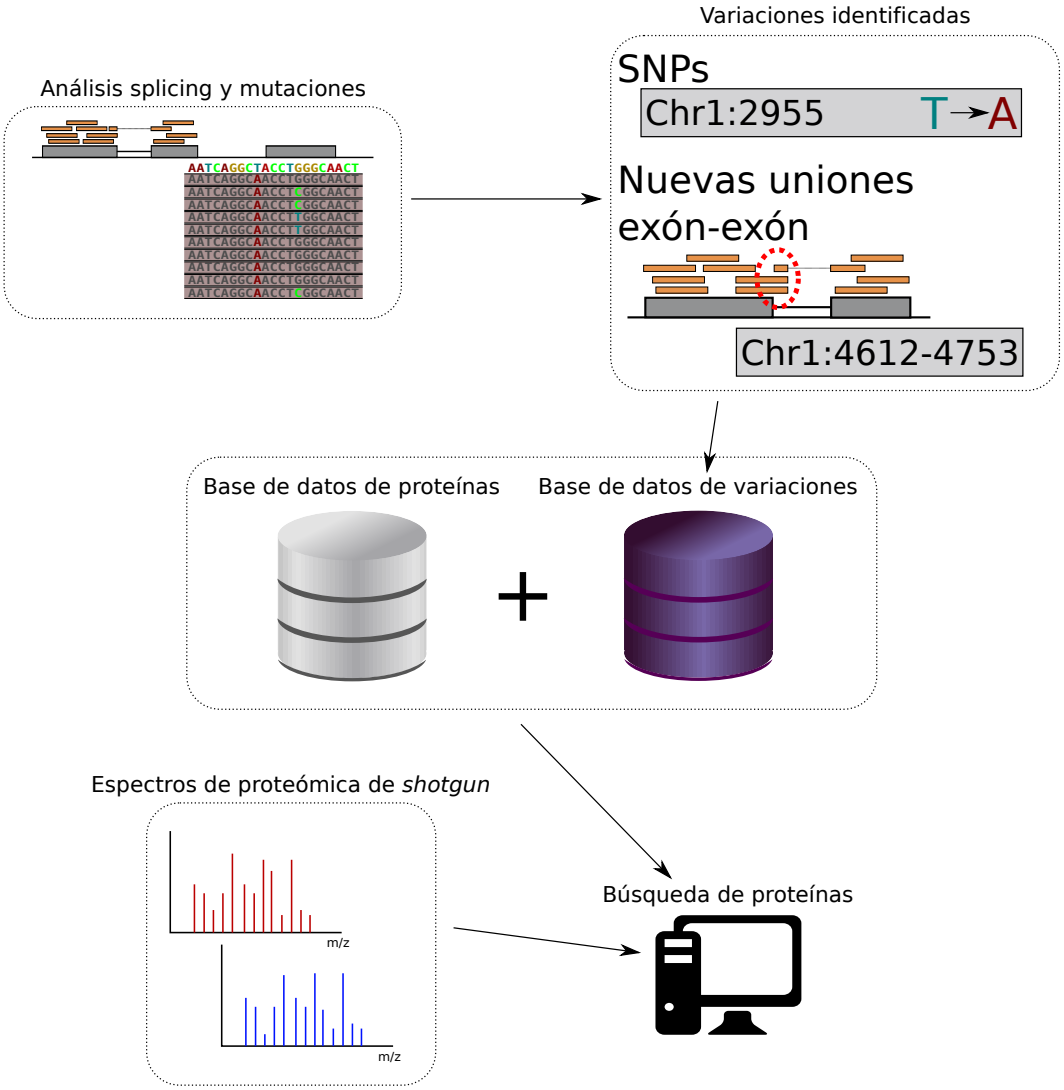


Figura 1.9: Esquema de flujo de trabajo de proteogenómica.

1.1.4.1.1 Procesamiento de *RNA-Seq*

1.1.4.1.1.1. Identificación de variantes El procesamiento a realizar no difiere mucho de un análisis de variantes estándar en un experimento de resecuenciación. Se parte de los ficheros *FASTQ* del secuenciador, que se alinean contra el fichero *FASTA* de referencia del organismo. Es importante utilizar un alineador con capacidad para reconocer el *splicing*, y que además pueda darnos un fichero de salida con las uniones entre exones detectadas al partir las lecturas. En el caso de los alineadores mencionados en la sección de transcriptómica, lo normal es obtener un fichero en formato *BED*, que no es más que un fichero de texto tabulado en el que podemos

encontrar las coordenadas de estas uniones de exones.

Posteriormente habría que realizar un análisis de variantes sobre el alineamiento, de la misma forma que el explicado en la Sección 1.1.2.2.3.2. De esta forma podemos obtener un fichero *VCF* que contenga todas las mutaciones y pequeñas inserciones y delecciones que se hayan detectado en nuestra muestra.

Estas diferencias con respecto al genoma de referencia son las que se recolectan para generar secuencias nuevas que permitan una mejor base para el proceso de búsqueda e identificación de proteínas.

1.1.4.1.1.2. Generación de bases de datos de péptidos Una vez tenemos toda la información posible sobre la variación de nuestra muestra a nivel transcriptómico, es necesario procesarla para generar nuevos péptidos que contengan esta variación. Existen varias formas de hacer esto, pero en análisis en los que el organismo está bien anotado, bastaría con generar nuevos péptidos en torno a los variantes encontrados. Hay que tener en cuenta los casos en los que al existir la mutación, pueda darse un cambio en el marco de lectura, ya que de esta forma no solo habría un cambio puntual en un aminoácido, sino que podría cambiar la secuencia de parte de la proteína. Existen programas como *customProDB*[300] o *PGTools*[207] que permiten realizar este trabajo partiendo de diferentes aproximaciones.

De esta forma obtenemos nuevas secuencias, que pueden ser almacenadas en un formato tipo *FASTA*. Conviene utilizar una notación suficientemente explicativa en las cabeceras de las secuencias, para poder identificar la proteína de origen, y la variación llevada a cabo sobre la misma.

1.1.4.1.2 Búsqueda de proteínas

Una vez que se han generado las nuevas secuencias procedentes de la variación detectada en el experimento transcriptómico, lo normal es concatenar el fichero *FASTA* generado con el procedente de las secuencias de proteínas de la referencia. Es posible realizar búsquedas separadas con las diferentes bases de datos, pero el porcentaje de falsos descubrimientos puede dispararse. Utilizando una base de datos de un tamaño similar al de la referencia original nos

permitirá posteriormente poder realizar comparaciones más robustas.

Existe también la posibilidad de filtrar las proteínas de la base de datos de referencia utilizando los datos de expresión procedentes del estudio transcriptómico. De esta manera puede mejorar la sensibilidad a la hora de encontrar proteínas que aparecen expresadas, pero con la contrapartida de que podemos perder la identificación de proteínas de las cuales no existe ARN mensajero.

Una vez decidida la base de datos a utilizar por el motor de búsqueda de proteínas, se realiza todo el procesamiento de igual manera, obteniendo identificaciones de péptidos que posteriormente se podrán inferir a proteínas. Ocurre en ocasiones que aparecen detectados nuevos péptidos que no son suficientes para determinar la existencia unívoca de una nueva forma de proteína, pero pueden dar pistas para la realización de nuevos experimentos.

1.1.5. Análisis terciario en ómicas

Una vez procesados los datos en experimentos de ómicas, en muchas ocasiones el resultado puede derivar en listas de genes o proteínas de interés para el experimento. Es tarea de la bioinformática investigar metodologías que permitan enriquecer los resultados de estos experimentos a la hora de mejorar la interpretación biológica de los mismos, o generar nuevas predicciones a partir de los mismos para poder guiar a los investigadores en la realización de nuevos experimentos. La información biológica disponible en repositorios y bases de datos públicas puede ser de gran utilidad en estos casos a la hora de realizar estas tareas.

1.1.5.1. Análisis de enriquecimiento funcional

Un análisis de enriquecimiento funcional consiste en, a partir de una lista de genes de interés procedente de un experimento, extraer información biológica sobre los mismos en forma de anotaciones provenientes de bases de datos y repositorios públicos, y entonces realizar un análisis estadístico para dilucidar cuales de esas anotaciones son relevantes a la hora de explicar esa lista. Existen multitud de métodos para realizar este tipo de análisis, donde lo más importante es un buen uso de las bases de datos biológicas para generar hipótesis lo más correctas

posible.

1.1.5.1.1 Repositorios biológicos

Este tipo de análisis aparecen de forma natural después de la creación de bases de datos de conocimiento biológico como *Gene Ontology*[16, 34]. *Gene Ontology* es una ontología en la cual se ha construido un vocabulario controlado que permite describir los genes y las características de sus productos. En realidad, está dividida en tres partes separadas, que engloban la totalidad de anotaciones para los genes, y son la función molecular de los productos de los genes, su participación en procesos biológicos, y su localización dentro de la célula. La ontología de *Gene Ontology* se ha implementado como un grafo acíclico dirigido, donde cada término se define como un nodo dentro del mismo, y las relaciones entre ellos son las aristas del grafo. Existe una cierta jerarquía dentro de estos términos, donde los nodos hijos son más especializados que sus padres. *Gene Ontology* no sólo se dedica a construir su ontología, sino que también realiza una anotación de genes con los términos de la misma. Esta anotación se lleva a cabo mediante dos metodologías distintas, una supervisada de asociación manual de contenidos, y otra mediante inferencia computacional. Gran parte de las herramientas de análisis de enriquecimiento de anotaciones parten de la utilización de esta ontología, por lo que en los años posteriores a su aparición (2002), aparecieron gran cantidad de métodos de este tipo. Quince años después, en 2017 *Gene Ontology* sigue siendo una referencia dentro del ámbito de este tipo de repositorios, y aún hoy en día sigue recibiendo financiación por parte de instituciones tan importantes como el *National Human Genome Research Institute (NHGRI)* perteneciente al *National Institutes of Health (NIH)*, por lo que se asegura la continuidad de este tipo de análisis.

Otras bases de datos interesantes para este tipo de estudios pueden ser las que estudian las rutas metabólicas de los organismos. Este es el caso de las bases de datos *KEGG*[304], *Panther*[197] o *Reactome*[107]. Estas bases de datos contienen una jerarquía de rutas metabólicas, y para cada una de ellas presentan los genes involucrados en las mismas, y la parte de la ruta en la que participan sus productos génicos.

Existen una gran cantidad de bases de datos que relacionan genes y proteínas con diversos tipos de información. Ejemplos de esto pueden ser por ejemplo *OMIM*[10] donde se relacionan genes con enfermedades mendelianas, *TRANSFAC*[181] donde se asocian genes con los factores de transcripción que los regulan, *miRBase*[143] que contiene información acerca de

micro ARNs e interacciones con genes o *PhosphoSitePlus*[117] que contiene modificaciones posttraduccionales de proteínas.

También hay una gran cantidad de bases de datos que proporcionan información a nivel de secuencia o estructura, como es el caso de *UniProt*[305], que contiene, a parte de una base de datos de proteínas de referencia con identificadores para las mismas, información acerca de la estructura secundaria y regiones de interés de las mismas, *Immune Epitope Database* (*IEDB*)[297] que contiene información acerca de epítopos de proteínas a nivel de secuencia, o bases de datos de mutaciones provenientes de diferentes enfermedades, tales como *dSysMap*[202] o *BioMuta*[310].

1.1.5.1.2 Tipos de análisis de enriquecimiento

Existe una gran cantidad de metodologías diferentes de análisis de enriquecimiento funcional. Según la revisión de *Huang et al.*[120], se pueden clasificar en tres tipos diferentes.

1.1.5.1.2.1. Análisis de enriquecimiento singular En este tipo de análisis, a partir de la lista de genes de interés, se va probando el enriquecimiento de cada una de las anotaciones asociadas a los genes de la lista una por una. Los términos que superen un umbral de enriquecimiento se reportan como la salida del análisis junto con los valores de este umbral. El enriquecimiento suele realizarse contando el número de genes de interés que están asociados a determinado término biológico, y comparar este número con el número global de genes asociados al mismo término en una lista de referencia para el experimento. Este cálculo se puede realizar con métodos estadísticos como Chi Cuadrado, el test exacto de *Fisher*, el cálculo de probabilidades en una distribución binomial o en una distribución hipergeométrica. La mayor parte de las herramientas que surgieron a partir del lanzamiento de *Gene Ontology* se basan en este tipo de análisis, que es el más sencillo de todos. Ejemplos de herramientas de este tipo de análisis son *DAVID*[121], *Onto-Express*[135] o *FATIGO+*[5]. Los resultados de estos análisis pueden ser listas de términos enormes, difíciles de interpretar, en los que no aparecen relaciones entre los mismos que puedan ayudar a los investigadores en la interpretación biológica de los mismos. El enriquecimiento de estos términos también puede variar dependiendo del tamaño de la lista seleccionado como entrada para este tipo de análisis, ya que no existen unas normas fijas a la hora de seleccionar los genes de la misma.

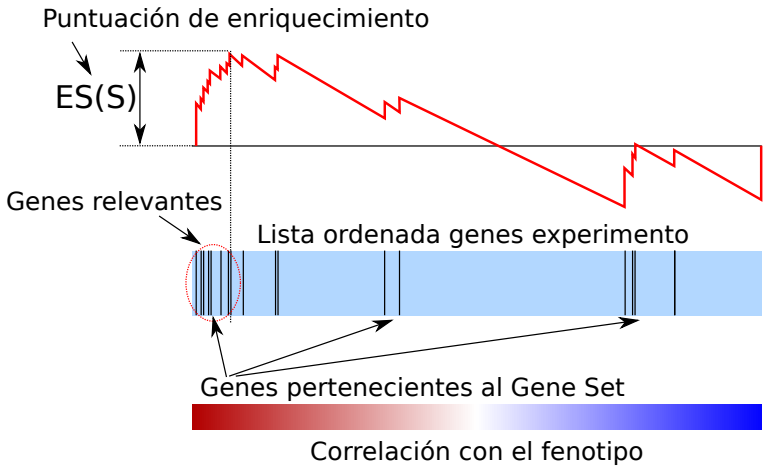


Figura 1.10: Esquema de resultados del método *GSEA*.

1.1.5.1.2.2. Análisis de enriquecimiento de conjuntos de genes Este método de enriquecimiento funcional, abreviado como *GSEA*[273], realiza una estrategia similar al análisis de enriquecimiento singular, pero en este caso no se utiliza una lista de genes de interés, sino que como entrada requiere la lista completa de genes con los valores asociados de expresión diferencial, ya que este tipo de análisis fue concebido para análisis de *microarrays*. Esto puede ser una ventaja, ya que genes que pueden ser descartados en un análisis singular, aquí pueden ser útiles contribuyendo al enriquecimiento de términos que de otra manera serían descartados. La estrategia en este método consiste en, a partir de esta lista completa de genes ordenada por los valores de expresión diferencial, encontrar la clasificación de los que pertenecen a una determinada categoría funcional, asignando un valor de enriquecimiento mayor cuantos más genes aparezcan en los extremos de la lista de genes ordenada, como puede observarse en la Figura 1.10. Esto puede llevarse a cabo con test estadísticos como el de *Kolmogorov-Smirnov*. Ejemplos de este tipo de herramientas son *FatiScan*[5] o *GO-Mapper*[264]. Un problema de este tipo de métodos es precisamente el hecho de que se tengan que utilizar listas de genes asociados a un valor que las ordene, ya que existen muchos tipos de análisis biológicos cuyo resultado pueden ser listas de genes, pero sin ningún valor de *ranking*. Un ejemplo de esto sería un análisis de resecuenciación de un genoma en busca de *SNPs*. La lista resultante de *SNPs* podría analizarse para generar una lista de genes mutados de interés, pero no existiría una forma directa de poder ordenarlos. Además, a pesar de tener en cuenta todos los genes de la lista ordenada para el análisis, los que aparecen en la zona central, es decir, los que tienen un valor mínimo de cambio de expresión, son mucho menos tenidos en cuenta por las estadísticas utilizadas, existiendo casos en los que alguno de estos genes puede llegar a ser relevante para el análisis.

1.1.5.1.2.3. Análisis de enriquecimiento modular El análisis de enriquecimiento modular toma como base el tipo de análisis utilizado en el singular, pero permite relacionar términos entre sí, por ejemplo agrupando términos compartidos por varios genes existentes en la lista de interés del experimento, pudiendo de esta forma mejorar la sensibilidad y especificidad del análisis. La forma de construir los grupos de anotaciones es importante, ya que utilizando datos de diferentes bases de datos biológicos, puede existir cierta redundancia en los mismos. La estadística utilizada en este tipo de análisis también es equivalente a la del análisis singular. Una de las limitaciones de este tipo de análisis es la posibilidad de que haya términos de los cuales no se encuentren relaciones entre ellos, y que queden fuera de los resultados significativos del análisis. Aún así, tener en cuenta las relaciones entre genes y anotaciones es algo mucho más cercano a la biología real del experimento. Ejemplos de herramientas de este tipo son *Ontologizer*[26] o *topGO*[7].

1.1.5.1.3 Listas de genes de referencia

En los análisis de enriquecimiento singular y modular, a la hora de realizar la estadística, es importante definir correctamente el conjunto de genes de referencia contra los que comparar al hacer el enriquecimiento. Diferentes listas de genes de referencia conducen a valores diferentes en la estadística, lo que conlleva p-valores de enriquecimiento distintos. Muchas herramientas utilizan por sistema el conjunto total de genes del genoma del organismo a analizar, o el total de genes del genoma que contienen anotaciones. Para análisis en los que la referencia del experimento sea todo el genoma, como por ejemplo en análisis de expresión diferencial utilizando datos de transcriptómica, esto puede ser correcto, pero en otros casos puede ser una elección inapropiada, como por ejemplo un experimento de *microarrays*, ya que los genes que no están presentes en el propio *microarray* no pueden tener la oportunidad de ser seleccionados para nuestra lista de interés. Por lo tanto, lo correcto es que la lista de referencia sea el conjunto de genes que pueden ser seleccionados para la categoría de anotaciones estudiada.

1.1.5.1.4 Corrección de múltiples hipótesis

En los análisis de enriquecimiento se prueban numerosas anotaciones de genes al mismo tiempo, por lo que la probabilidad de que aparezcan falsos positivos aumenta[27]. Se debe realizar una corrección de comparaciones múltiples debido a esto. Muchas de las herramientas

existentes utilizan métodos como el de *Bonferroni* o *Holm*. Estos métodos asumen independencia de las variables, pero en este caso pueden existir dependencias entre las mismas, por lo que métodos como el de *Benjamini-Hochberg* pueden ser más apropiados. De todas formas, los p-valores de enriquecimiento son muy dependientes de los métodos utilizados, y estos tests pueden ser muy conservadores realizando la corrección, pudiendo eliminar anotaciones de los resultados que pueden llegar a ser relevantes para la correcta interpretación del análisis. Por lo tanto, es importante conocer las limitaciones de estos métodos, y saber que un mejor análisis puede deberse sobre todo por la utilización de mejores fuentes de datos (anotaciones, mapeo de genes).

1.1.5.1.5 Traducción de identificadores

Existe una gran variedad de identificadores de genes y proteínas distintos mantenidos por organizaciones independientes, y que generan una gran redundancia. A la hora de relacionar anotaciones biológicas con genes, los diseñadores de estas bases de datos deben tomar la decisión de cuál de estos identificadores utilizar. Poder traducir entre diferentes tipos de identificadores de genes y proteínas y hacerlo de una manera eficiente es una característica muy necesaria en las herramientas de enriquecimiento. De no ser así, los análisis podrían quedar sesgados, y no explicar de manera correcta la biología del experimento. También es importante poder traducir identificadores de sondas, como las de los *microarrays*, ya que muchos análisis de este tipo provienen de resultados de expresión diferencial con este tipo de tecnología, y poder integrar todos los pasos de traducción de identificadores en una misma herramienta minimiza el número de errores cometidos en el proceso. Gran parte de las herramientas de análisis toman como referencia algún tipo de identificadores de genes, como *Entrez Gene*[173] o *Ensembl*[124]. Existen también esfuerzos por crear sitios centralizados para obtener información de diversas fuentes de datos de genes, incluyendo traducciones entre identificadores, como es el caso de *Biomart*[263].

1.1.5.2. Análisis de perfiles de expresión

Los resultados de experimentos de transcriptómica suelen contener la lista de genes del experimento (en *microarrays* los genes asociados a las sondas presentes en el mismo, en *RNA-Seq* la lista completa de genes del organismo secuenciado) con un valor numérico asociado, que puede ser un valor de expresión crudo o normalizado, o un valor de diferencia de expresión

en el caso de comparar dos condiciones. En la actualidad, suele ser obligatorio, o al menos recomendable, a la hora de publicar un artículo de investigación donde se han realizado análisis de este tipo, publicar estos datos de expresión en bases de datos públicas, tales como *Gene Expression Omnibus* (*GEO*)[24] o *ArrayExpress*[141] entre otras[248]. Estas bases de datos almacenan, junto con estos datos de expresión, datos acerca del experimento realizado, incluyendo información acerca de la preparación de las muestras, las características de la muestra, como el tipo de tejido o cultivo celular del que procede, o las condiciones comparadas. Poder extraer información de los experimentos presentes en estas bases de datos nos permite sobre todo poder realizar meta-análisis comparando diferentes condiciones, para poder extraer conclusiones valiosas sobre datos ya publicados, que de otro modo no podrían reaprovecharse. Existen numerosos ejemplos de meta-análisis publicados en revistas de investigación de alto impacto[309, 172]. El tipo de meta-análisis que se puede realizar depende en gran medida de la información almacenada asociada a los experimentos, ya que gracias a ella se pueden agrupar diferentes experimentos.

1.1.5.2.1 Bases de datos de experimentos

GEO y *ArrayExpress* son las bases de datos de experimentos más populares en la actualidad. De hecho existe información duplicada en ambas bases de datos, ya que *ArrayExpress* importa sistemáticamente parte de los datos publicados en *GEO*[91]. También existe información redundante dentro de cada una de estas bases de datos debido entre otros a la reutilización de datos de experimentos en diferentes publicaciones[244]. Es importante conocer la estructura y tipos de datos existentes en estas bases de datos para poder realizar búsquedas y análisis coherentes en ellas. Existen otras librerías de perfiles de expresión que utilizan cultivos celulares u organismos modelo para poder realizar también meta-análisis sobre ellos, como es el caso de el programa *LINCS* del *NIH*, o *Connectivity Map* (*CMAP*)[149] del *Broad Institute*.

1.1.5.2.1.1. *Gene Expression Omnibus* *GEO* es una base de datos pública perteneciente al *NCBI* que almacena datos procedentes de *microarrays*, experimentos de *NGS* y otros tipos de datos procedentes de genómica funcional y que son enviados por la comunidad investigadora. Los datos almacenados en *GEO* cumplen con el estándar *MIAME*[37], que son básicamente unas guías y requisitos sobre la información mínima que debería ser almacenada en un experimento de *microarrays*. Más tarde, para poder empezar a incluir datos de *NGS* en *GEO*, se

adoptó también el estándar *MINSEQE*[79]. En *GEO* existen principalmente dos tipos de datos, los que se almacenan en el mismo formato en el que los enviaron los autores, y conjuntos de datos procesados. La organización de estos tipos de datos puede visualizarse en la Figura 1.11.

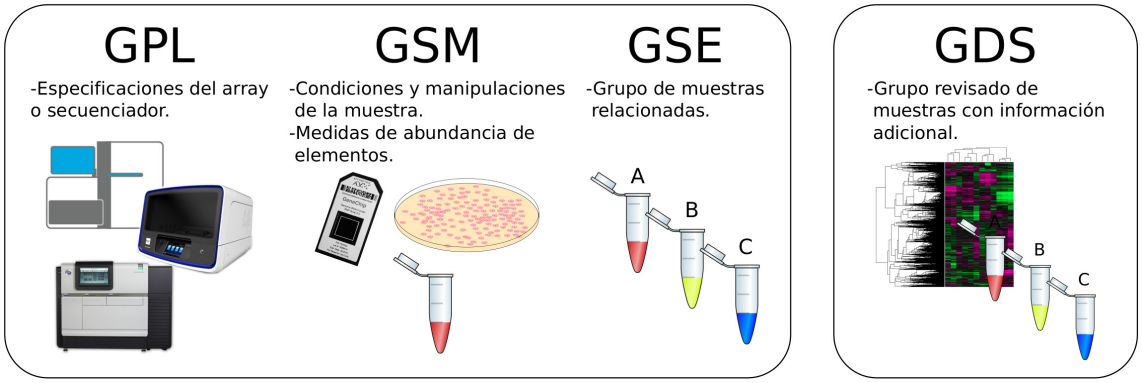


Figura 1.11: Esquema de organización de los datos en la plataforma *GEO*.

Para los datos enviados por los autores existe una organización para regular y estandarizar el formato de datos. Las *GPL* o plataformas son una descripción de las características del aparato utilizado para realizar el análisis, como por ejemplo el tipo de *microarray*, o el secuenciador utilizado, además de una plantilla con las sondas o genes utilizados para el mismo. De esta forma, con un identificador de plataforma, podremos saber para qué identificadores de gen podremos tener valores de expresión. Las *GSM* o muestras, describen por un lado los protocolos que se han seguido para manipular la muestra y analizarla, y las medidas de abundancia para cada una de las sondas o genes presentes en la plataforma que tiene asociada. Por último están las *GSE* o series, que son agrupaciones de muestras que están relacionadas entre sí en un experimento. También pueden incluir datos o análisis adicionales que no tienen cabida en otro sitio. Existe gente en *GEO*, que a partir de toda esta información, construye nuevos conjuntos de datos llamados *GDS*. Un *GDS* es un conjunto de muestras *GSM* que son biológica y estadísticamente comparables entre sí. Muestras de un mismo *GDS* deben pertenecer además a la misma plataforma *GPL*, por lo que comparten un mismo conjunto de identificadores, cuyos valores de expresión también deben de estar calculados de una forma equivalente, utilizando por ejemplo un mismo método de normalización. Las muestras pertenecientes a un *GDS* también contienen información en forma de etiquetas sobre los factores experimentales, pudiéndose subdividir los *GDS* por grupos con etiquetas comunes. *GEO* contiene en su portal herramientas para la búsqueda de estos datos, su comparación, y visualizaciones de los mismos.

1.1.5.2.1.2. ArrayExpress *ArrayExpress* es un repositorio de experimentos de genómica funcional muy similar a *GEO*, pero perteneciente al *EMBL-EBI*. Al igual que *GEO*, el repositorio contiene principalmente datos de *microarrays* y experimentos *NGS* que cumplen con los estándares de *MIAME* y *MINSEQE* respectivamente. Las fuentes de datos de *ArrayExpress* son principalmente dos, los datos enviados por los investigadores, que posteriormente son supervisados y refinados por trabajadores del *EMBL-EBI*, y la importación de datos de *GEO*. La organización de datos también es similar a la de *GEO*. En este repositorio los datos van organizados en torno a experimentos, que pueden contener varios ensayos que pertenecen a un estudio. Los experimentos almacenan metadatos describiendo la muestra biológica y el procedimiento experimental, así como tablas de resultados. En los experimentos de *microarrays* se almacenan los ficheros de resultados en crudo, y resultados procesados en forma de matrices de expresión. Para experimentos *NGS*, son almacenados los ficheros *BAM* de alineamientos y ficheros de resultado en forma de matrices con datos de expresión normalizados. Los ficheros de secuencias en crudo se depositan en otro repositorio llamado *European Nucleotide Archive* (*ENA*). Para las diferentes muestras pertenecientes a los ensayos en este repositorio además se almacenan unos metadatos llamados variables experimentales, que son básicamente atributos de las muestras que describen factores que difieren entre las muestras de prueba y las de control. Esta información también es muy similar a la existente en las etiquetas de los *GDE* de *GEO*. En esta plataforma también existe una herramienta de búsqueda para poder acceder fácilmente a experimentos a través de sus metadatos.

1.1.5.2.1.3. *Connectivity Map* y *LINCS* *Connectivity Map* se construyó con el propósito de poder analizar librerías de compuestos químicos y fármacos sobre una gran diversidad de tipos de tejidos y líneas celulares humanas, y poder descubrir conexiones entre diferentes compuestos utilizando los perfiles de expresión resultantes. Contiene más de un millón y medio de perfiles de expresión generados a partir de ensayos utilizando *microarrays* en los que se han utilizado unos 5000 compuestos químicos y 3000 reactivos genéticos. Los perfiles de expresión asociados a la adición de un compuesto contienen una serie de metadatos tales como el tipo de tejido o línea celular utilizada, la concentración y duración del tratamiento con el compuesto químico. También contiene herramientas de búsqueda a partir de estos metadatos, y de comparación de perfiles de expresión, utilizando algoritmos de reconocimiento de patrones simples muy parecidos a los utilizados en la herramienta *GSEA* de enriquecimiento de anotaciones para poder encontrar estas asociaciones entre perfiles y compuestos.

LINCS es una librería de perfiles de expresión muy similar a *CMAP*, que contiene conjuntos de datos provenientes de resultados de ensayos utilizando líneas celulares y tejidos de humanos con diferentes compuestos. Los perfiles generados en este repositorio son en su mayoría perfiles de expresión de subconjuntos de genes que sean representantes de todo el transcriptoma. Este parecido con *CMAP* ha llevado a que se hayan aunado fuerzas y se haya creado la infraestructura de cómputo *CLUE* (*CMAP and LINCS Unified Environment*)[51], que es un conjunto de herramientas y aplicaciones web que permita acceder y manipular a los usuarios a todo este conjunto de datos de una forma unificada.

1.1.5.2.2 Análisis y comparación de perfiles

Con el auge de la tecnología de los *microarrays* se extendió su uso para realizar comparaciones entre diferentes muestras y poder extraer conclusiones biológicas midiendo las diferencias y similitudes entre ellas. La forma más sencilla de comparación sería el análisis de expresión génica explicado anteriormente en los flujos de trabajo de transcriptómica, donde las muestras se comparan dos a dos, y se mide la diferencia de expresión a nivel de genes. Un paso más allá es el estudio de matrices de expresión génica, que son matrices generadas a partir de la expresión génica de un número mayor de muestras. A este nivel se pueden comparar las diferencias entre las muestras o entre genes, simplemente comparando filas o columnas de la matriz. Para poder realizar estas comparaciones es necesario poder definir métricas de distancias entre muestras o genes (filas o columnas). Al tener vectores numéricos con la expresión génica para cada gen para cada muestra, una posibilidad sería recurrir a la distancia euclídea o a la de correlación de *Pearson*, pero existen una gran cantidad de métricas que pueden ser más adecuadas para este propósito, como las recogidas en [128]. Una vez elegida una métrica adecuada, hay que decidir si utilizar información adicional sobre las muestras o genes presentes en la matriz de expresión, ya que dependiendo de esto podremos hacer un tipo de análisis u otro.

Cuando no disponemos de información adicional sobre las muestras podemos optar por intentar clasificarlas mediante algoritmos de *clustering* para así agrupar grupos de genes o muestras con propiedades características. Existen numerosas técnicas de *clustering* no enfocadas en expresión génica pero que se utilizan en este campo con éxito, tales como el *clustering* jerárquico[70], *k-means*[236], mapas autoorganizados[277], análisis de componentes principales[130], la descomposición en valores singulares[93] e incluso la factorización no negativa de matrices[218, 152].

Al disponer de información adicional sobre las muestras, una posibilidad sería la de construir clasificadores tales como discriminantes lineales, árboles de decisión o máquinas de vectores soporte que permitan asignar clases predefinidas a perfiles de expresión génica dados como entrada. Normalmente se suelen entrenar estos clasificadores con un conjunto de datos conocido, para posteriormente utilizarlos para ayudar a distinguir entre, por ejemplo, subtipos conocidos de tipos de cáncer[99].

En procesos exploratorios que implican una gran cantidad de muestras, como las presentes en las bases de dato explicadas anteriormente, realizar alguna de las aproximaciones de *clustering* anteriormente mencionadas puede ser inabarcable por la complejidad computacional y el tamaño de las mismas, así como construir clasificadores sobre conjuntos tan heterogéneos de muestras, que puede no ser demasiado útil, ya que estos se suelen centrar en intentar clasificar perfiles de expresión con unas características determinadas. En estos casos, una opción podría ser la de intentar buscar conexiones entre la expresión de una muestra de entrada, y los perfiles almacenados en la base de datos. De esta forma podemos relacionar los metadatos almacenados en estos perfiles similares con la muestra de entrada. Unas primeras aproximaciones en este sentido [125, 239] en las cuales se utilizaban conjuntos más pequeños y específicos de datos, dieron paso a aproximaciones más sofisticadas como la previamente mencionada de comparación de perfiles utilizando un algoritmo parecido al de *GSEA* en *Connectivity Map*[149] o *MARQ*[291], que con una estrategia similar permitía comparar perfiles con un subconjunto amplio de datos escogidos de *GEO*.

1.1.5.3. Predicción de interacciones en elementos regulatorios

Uno de los retos de la biología molecular es el de entender los mecanismos de regulación de la expresión génica. Se han descubierto gran cantidad de mecanismos para llevar a cabo esta regulación: modificaciones en el estado de la cromatina, mediante, por ejemplo, modificaciones en las histonas, que pueden hacer que no esté disponible para las enzimas de transcripción el acceso a las regiones del ADN necesarias para que se inicie el proceso; la metilación del ADN que permite silenciar genes o regiones enteras de ADN; el ARN no codificante, cuyas secuencias son complementarias al ADN o ARN codificante e impiden su traducción. Existen diferentes tipos de ARN no codificante, tales como el ARN pequeño nuclear, el ARN pequeño de interferencia, ARNs asociados a Piwi, o los micro ARNs. Estos últimos han sido foco de atención en estos últimos años debido a su participación en la regulación de gran cantidad de

procesos de la celula.

1.1.5.3.1 Micro ARNs

Los micro ARNs[154] se encuentran entre los tipos existentes de ARN no codificantes, habiendo adquirido una gran relevancia desde su descubrimiento en 1993, ya que se ha asociado a su regulación con procesos tales como la apoptosis[46], proliferación celular[38] y otros procesos fisiológicos y patológicos[102]. Los micro ARNs son unas moléculas de ARN monocatenario de aproximadamente 22 nucleótidos que se unen a un complejo proteico denominado *RISC*, y lo guían a ARNs mensajeros de los cuales son complementarios en secuencia, tal y como puede apreciarse en la Figura 1.12. Estas moléculas existen en gran cantidad de organismos, existiendo ligeras diferencias en su comportamiento entre ellos. En plantas por ejemplo, los micro ARNs tienen una complementariedad perfecta con las regiones codificantes de sus dianas, produciendo la degradación del ARN mensajero[25]. En animales generalmente se unen de una manera imperfecta a la región *3'UTR* del mensajero produciendo una inhibición en la traducción del mismo[40].

A la hora de poder predecir interacciones de sitios de unión entre micro ARNs y ARN mensajeros, es necesario conocer cómo funciona la unión entre estas dos moléculas. En mamíferos por ejemplo existen varias reglas definidas experimentalmente. La complementariedad de secuencia entre la semilla del micro ARN, que es una parte del mismo que normalmente se ubica en los nucleótidos entre el 2 y el 7, y el mensajero, suele ser suficiente para producir la represión del ARN mensajero[38, 159]. La complementariedad de secuencia en los nucleótidos del 13 al 16 del micro ARN puede reforzar la afinidad con el mensajero, o incluso compensar una complementariedad incompleta con la semilla. Los emparejamientos *wobble* entre nucleótidos G y U en la región de la semilla puede interferir en la unión con el ARN mensajero[64]. La conformación espacial del mensajero, y la estabilidad termodinámica en la unión con el micro ARN también son muy importantes a la hora de que se produzca la unión[134].

MirBase[144] es la base de datos más importante de información acerca de micro ARNs, conteniendo un repositorio de secuencias de estas moléculas en su forma tanto madura como precursora. En los últimos años el conocimiento experimental acerca de estas moléculas ha crecido en gran medida gracias a experimentos de *NGS* tales como el *HITS-CLIP*[47], en los que se inmunoprecipitan proteínas del complejo *RISC* pudiendo de esta forma identificar ARN

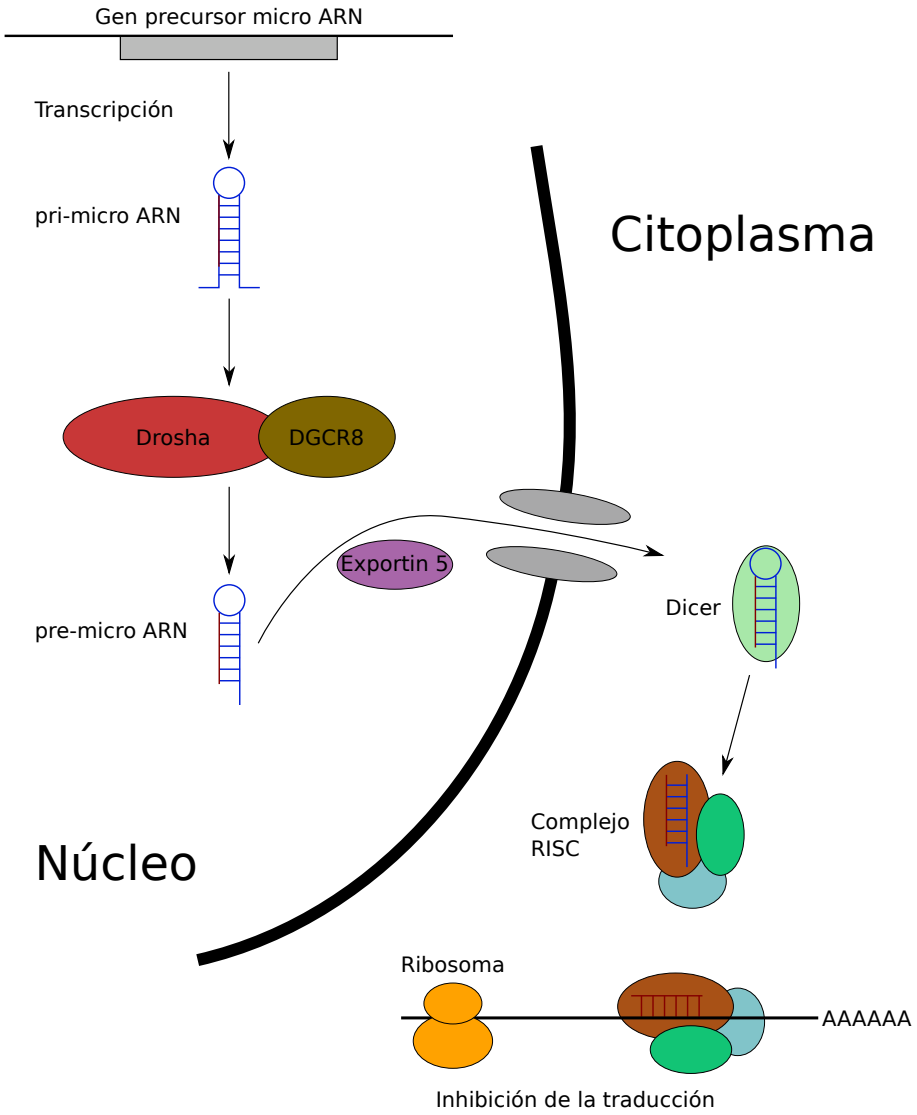


Figura 1.12: Esquema del proceso de silenciamiento de genes por micro ARNs.

mensajeros dianas de determinados micro ARNs.

Existen también gran cantidad de bases de datos de interacciones validadas experimentalmente. Ejemplos de este tipo serían *TarBase*[258, 223], *miRTarBase*[118, 119, 48] o *starBase*[314, 162]. Estas bases de datos suelen incluir interacciones de varios organismos, y en algunos casos organizadas por el tipo de experimento por el cual se ha validado la interacción, existiendo evidencias más o menos fiables. La base de datos *starBase* es algo más novedosa, ya que utiliza las técnicas previamente mencionadas de *HITS-CLIP* y similares del ámbito del *NGS* para aumentar de forma significativa la cantidad de datos validados de interacciones.

1.1.5.3.1.1. Predicción de interacciones Conocer todas las posibles interacciones entre micro ARNs y ARN mensajero es crucial para entender ciertos aspectos de la regulación de la célula. Gracias al trabajo de gran cantidad de investigadores el tamaño de las bases de datos validadas experimentalmente no para de crecer. Pero aún así, sigue habiendo una gran cantidad de posibles interacciones que no han sido todavía estudiadas, por lo que también se está llevando a cabo un gran esfuerzo en el desarrollo de algoritmos de predicción de interacciones. Los métodos existentes de predicción pueden ser de varios tipos[238]: *ab initio*, aprendizaje automático (*machine learning*) y métodos híbridos.

Los algoritmos *ab initio* se basan en las reglas conocidas estudiadas en las interacciones validadas experimentalmente. Se basan en modelos computacionales que no utilizan los datos experimentales directamente. Por ejemplo, *MiRanda*[74] utiliza una puntuación de complementariedad estimada para seleccionar a las parejas. *MicroTar*[280] y *PITA*[134] buscan diferentes tipos de complementariedad dentro de la semilla, permitiendo además *wobbles*. Otros algoritmos como el de *TargetScan*[159] primero buscan una complementariedad perfecta dentro de la semilla, y luego generan una puntuación basándose en diferentes aspectos del contexto de la unión. Casi todos estos métodos utilizan el paquete *Vienna RNA*[113] para calcular la estabilidad termodinámica y utilizarla como parte de la puntuación final de la interacción, e incluso algoritmos como el de *RNAhybrid*[237] o *miRiam*[148] primero intentan maximizar esta estabilidad antes de buscar la complementariedad entre las secuencias.

Los algoritmos de *machine learning* en cambio utilizan los datos experimentales para poder entrenar un clasificador. De esta forma el clasificador puede identificar sitios diana basándose en en la similitud con sitios ya conocidos. Como ejemplo de este tipo de algoritmos está *RFMirTar-*

get[192], que utiliza un clasificador *random forest* que evalúa 17 características de predicciones realizadas por el algoritmo *miRanda* en un conjunto de datos de prueba. En el caso de *MultiMiTar*[198], el clasificador es una máquina de vectores soporte que utiliza 90 características de las interacciones. Para estos programas, se utilizaron como ejemplos de interacciones positivas las pertenecientes a las bases de datos experimentales *miRecords* y *TarBase*.

Los métodos híbridos aparecieron debido a problemas existentes en los algoritmos *ab initio*. Su problema es la alta tasa de falsos positivos que contienen, o en algunos casos, el excesivo filtrado para intentar aplacar esta alta tasa. Las aproximaciones híbridas son básicamente algoritmos *ab initio* a los cuales posteriormente se les aplica clasificadores entrenados con características extraídas de bases de datos experimentales. De esta forma, el número de falsos positivos se reduce. Por otro lado, el problema de las aproximaciones de aprendizaje automático es la poca cantidad de información acerca de interacciones de las cuales se ha validado su no existencia. Como ejemplo de aproximación híbrida está *NBmiRTar*[316], que primero aplica el algoritmo *ab initio miRanda*, y después utiliza un clasificador bayesiano ingenuo para filtrar la salida de éste. Para el clasificador utiliza datos de interacciones experimentales existentes en *TarBase*.

Muchos de estos algoritmos que generan predicciones de interacciones han incorporado sus resultados en bases de datos tales como *EIMMo*[85], *DIANA-microT*[175], *TargetScan*[158] o *PITA*[134].

1.1.5.3.1.2. Combinación de predicciones Debido a los problemas existentes con los algoritmos *ab initio* y basados en predictores de aprendizaje automático, también se han propuesto métodos de combinación de diferentes métodos y bases de datos para intentar mejorar la sensibilidad y especificidad de las predicciones. Por ejemplo, en [258] se midió la capacidad de varios algoritmos y bases de datos predictivas, así como combinaciones entre ellas. La conclusión de este estudio fue que el valor más alto de especificidad se logró con la intersección de cinco de estos algoritmos. En el caso de *ComiR*[53], se combinan cuatro bases de datos estimando la probabilidad de cada gen de ser diana de un conjunto de entrada de micro ARNs utilizando un algoritmo de máquinas de vector soporte, considerando además, en el caso de que se proporcione, valores de expresión de estos ARN. *ExprTarget*[86] utiliza una regresión logística para combinar las puntuaciones de predicciones en diferentes bases de datos con datos de expresión para ARN mensajero y micro ARNs. Este algoritmo utiliza estos valores de expresión para ajus-

tar un modelo lineal para cada posible interacción y el p-valor obtenido se utiliza como parte de la puntuación en el modelo. Además las puntuaciones de predicciones de diferentes bases de datos se ajustan en función de su parecido con los valores de bases de datos experimentales. *BCmicrO*[317] utiliza un modelo probabilístico para determinar la probabilidad de una interacción de ser experimentalmente validada utilizando puntuaciones de diferentes bases de datos. Existen muchos más ejemplos de algoritmos de este tipo que se basan en combinar puntuaciones de diferentes bases de datos predictivas, como por ejemplo *Ranking Aggregation*[60], *GenMiR3*[122] o *a Bayesian Graphical model*[270].

Capítulo 2

Objetivos

El desarrollo de nuevas técnicas experimentales de alto rendimiento en biología, tales como el *High Throughput Sequencing* y la proteómica de *shotgun*, están permitiendo un conocimiento mucho más profundo de los mecanismos de funcionamiento celulares. Pero esto ha tenido como consecuencia la generación de resultados cada vez más complejos y de mayor tamaño, difíciles de analizar e interpretar por los especialistas sin el apoyo de nuevas herramientas informáticas. Es necesario crear herramientas que sean lo suficientemente eficientes para manejar este tipo de datos en un tiempo razonable, por lo que cada vez más se están empezando a incluir técnicas de computación de altas prestaciones que incorporen el uso de granjas de computación, computación paralela y gestión de plataformas virtualizadas.

Los objetivos generales de esta tesis incluyen un abordaje integral del análisis masivo de datos provenientes de los campos de la transcriptómica y proteómica, mediante el desarrollo de nuevos métodos y flujos de trabajo que abarquen desde el procesamiento de los datos en crudo hasta la obtención de resultados de alto nivel que puedan ser enriquecidos con información de bases de datos y ontologías de libre acceso, y de esta forma construir aplicaciones que puedan facilitar el análisis de los datos por parte de los biólogos experimentalistas, reduciendo el tiempo de procesado y generando visualizaciones e informes que permitan una fácil interpretación de los mismos y crear nuevas hipótesis que sirvan de base para el desarrollo de nuevos experimentos de validación.

De forma más específica, se proponen los siguientes objetivos:

1. Diseñar un flujo automático de procesamiento de datos de *RNA-Seq* que implique un análisis de calidad y procesamiento primario de los datos, hasta llegar a información en forma de cuantificación de genes, diferencias de expresión entre condiciones experimentales, mutaciones y nuevos tránscritos. Abordar la normalización de la expresión génica y el cálculo de umbrales de expresión para los genes, así como el desarrollo de métodos proteogenómicos útiles para generar nuevas búsquedas con datos de proteómica de *shotgun*. Enmarcar este desarrollo dentro de la utilización de repositorios de experimentos de acceso público, para poder analizar y comparar una gran cantidad de muestras de forma simultánea.

2. Desarrollar herramientas de análisis terciario a partir de los resultados de alto nivel del método anterior. Por un lado, utilizando toda la información biológica disponible en bases de datos públicas en forma de términos biológicos asociados a genes, agrupar conjuntos de genes provenientes de estos resultados en base a estos términos para averiguar los procesos en los que estos genes están implicados. Por otro lado, agrupar, anotar y procesar datos de resultados de expresión de repositorios públicos para poder compararlos con nuevos experimentos, y realizar asociaciones en base a las anotaciones asociadas a los mismos.

3. Desarrollar metodologías para realizar predicciones de elementos regulatorios de la expresión de genes como los micro ARNs, integrando características de diferentes métodos ya existentes para maximizar la probabilidad de acierto a la hora de realizar una validación experimental. Generar además nuevos métodos de comparación entre diferentes algoritmos de predicción que lo hagan de una manera más precisa y tolerante a la falta de información que los métodos existentes.

4. Desarrollar una herramienta de integración de datos de biología estructural con información biológica a nivel de secuencia procedente de bases de datos públicas, con la posibilidad de ahondar más en el conocimiento de determinados genes o proteínas resultantes de los análisis previos.

En todos estos casos se utilizarán métodos de computación de altas prestaciones para optimizar los tiempos de ejecución de estas herramientas. Además, utilizando tecnologías web, crear visualizaciones interactivas para estas herramientas para facilitar al máximo la interpretación de estos resultados.

Capítulo 3

Aportaciones principales

3.1. *Proteogenomics Dashboard for the Human Proteome Project* (*dasHPPboard*)

dasHPPboard es un novedoso panel de datos orientado a la proteómica, en el que se recogen datos tanto en crudo como procesados pertenecientes al consorcio español del proyecto del proteoma humano (*SpHPP*), con el objetivo de ayudar al *HPP* a completar el mapa del proteoma humano. El *HPP* es un proyecto perteneciente a la *Human Proteome Organization* (*HUPO*), cuya finalidad es la de encontrar y caracterizar todas las proteínas codificadas por genes humanos para de esta forma generar un mapa de la arquitectura molecular del cuerpo humano basado en proteínas y ayudar a mejorar en el conocimiento, diagnóstico y tratamiento de enfermedades. Este proyecto está basado en tres pilares: espectrometría de masas, anticuerpos y reactivos de captura por afinidad, y bases de conocimiento fundamentadas en la bioinformática[155]. Desde su inicio, el *HUPO* dividió el proyecto *HPP* en dos programas: *C-HPP*[220] que basa el trabajo en la caracterización del proteoma haciendo una división por cromosomas, y *B/D-HPP*[2], que divide el trabajo en grupos por diferentes enfermedades y sistemas biológicos. En este proyecto es de gran importancia el estudio de las proteínas perdidas (*missing*), que son aquellas de las cuales sólo se tiene evidencia a nivel de tránscrito y una secuencia predicha (o inferida por homología) o proteínas parcialmente identificadas de las cuales hay evidencia transcriptómica pero sin información convincente proveniente de espectrometría de masas. De hecho constituye

la primera fase del *C-HPP* y se espera que concluya en 2018[219]. Dentro de esta división, el *SpHPP* se encuentra dentro del proyecto *C-HPP*, siendo responsable del cromosoma 16. El consorcio español a su vez está dividido en cinco grupos de trabajo: expresión de proteínas y estándares de péptidos, plataforma *SRM*, cuidado de salud clínico y biobanco, secuenciación de proteínas, y bioinformática[255]. En el cromosoma 16, según *neXtProt* (versión 2017-1-23)[87] existen 854 genes que codifican proteínas, de los cuales 82 son *missing*.

El desarrollo de *dasHPPboard* se enmarca en la pertenencia durante el año 2014 al grupo de trabajo de bioinformática del *SpHPP*. Debido a la gran cantidad de tipos de células diferentes en el cuerpo humano, existe gran variabilidad en la expresión de los genes y proteínas. No todos los genes y proteínas se expresan en todos los tipos de tejidos, por lo que se deben utilizar muestras de diferentes tejidos y líneas celulares para mejorar la eficiencia de la búsqueda de estas proteínas *missing*[101]. A pesar de esto, muchos grupos de investigación siguen utilizando líneas celulares y tejidos por su disponibilidad, o facilidad de utilización. Gracias a la existencia de estudios transcriptómicos que contienen gran cantidad de datos, como el proyecto *ENCODE*[206], que contiene una gran cantidad de datos de experimentos de líneas celulares humanas, se puede realizar una mejor selección de las muestras buscando evidencias transcriptómicas de la expresión de determinados genes antes de iniciar experimentos proteómicos. Existen varias herramientas para buscar proteínas dentro de experimentos proteómicos (*the Proteome Browser*[96], *CAPER*[100], *GenomewidePDB*[129], *H-InvDB*[313], *Chromosome 18 Knowledgebase*[231]), pero *dasHPPboard*, aparte de ser un panel de visualización y búsqueda de experimentos transcriptómicos y proteómicos, además incluye las bases de datos necesarias para poder llevar a cabo estudios proteogenómicos, gracias a la detección de polimorfismos y nuevas uniones entre exones.

3.1.1. Bases de datos de experimentos

El valor del *dasHPPboard* reside principalmente en la cantidad de datos de diferentes tipos de experimentos tanto de proteómica como de expresión génica que contiene. Como se ha dicho en los parrafos anteriores, una búsqueda efectiva de proteínas *missing* requiere de una primera fase de descubrimiento de tejidos y líneas celulares en los que estas proteínas se encuentran expresadas. La existencia de grandes bases de datos públicas de experimentos de expresión génica son de gran utilidad en este sentido, ya que la expresión génica es un indicador de la expresión proteica, habiéndose integrado datos de *ENCODE*, y datos procedentes de *GEO* o

ArrayExpress. También se han incluido datos proteómicos del *SpHPP*, ya que esta herramienta se pensó inicialmente como un repositorio de experimentos del *HPP*. La integración de estos tipos de datos a través de la proteogenómica es otro de los grandes pilares de la herramienta.

3.1.1.1. *SpHPP*

Uno de los objetivos de este trabajo consistía en el desarrollo de una base de datos cuyo cometido era el de contener todos los datos producidos por el proyecto *HPP*. En una primera aproximación se han integrado los datos provenientes del *SpHPP*. El *SpHPP*, como responsable del cromosoma 16 dentro del proyecto *C-HPP*, inicialmente centró sus esfuerzos en la realización de experimentos proteómicos de espectrometría de masas utilizando líneas celulares conocidas para intentar validar experimentalmente la presencia de proteínas *missing*.

Estos incluyen experimentos de proteómica de *shotgun*, utilizando espectrómetros de masas de alto rendimiento (*Orbitrap*, *Q Exactive*, *MaXis Impact* y *5600 triple TOF*), para caracterizar las líneas celulares *MCF7*, *CCD18*, *Jurkat* y *Ramos*. Estos datos de espectros y búsquedas se encuentran almacenados en la base de datos de *proteomeXchange*[298] en las accesiones *PXD000442*, *PXD000443*, *PXD000447* y *PXD000449*.

También se han integrado datos de *SRM (Selected Reaction Monitoring)* generados en los laboratorios de proteómica del *SpHPP*. Este tipo de técnica permite utilizar la espectrometría de masas en tándem para buscar y cuantificar una o unas pocas proteínas en una muestra compleja, por lo que se suele utilizar como validación una vez que existen datos positivos en un experimento de *shotgun*. Estos experimentos *SRM* se realizaron utilizando al menos tres líneas celulares utilizadas por todos los laboratorios del *SpHPP*, *MCF7*, *CCD18* y *Ramos*. También se han integrado datos de otros experimentos realizados con las líneas celulares *TC28*, *Jurkat* y *HUH7*, y con tejidos como celulas humanas endoteliales de la arteria umbilical (*HUAEC*), plasma y suero. En total, los resultados con estos tejidos y líneas celulares incluyen métodos *SRM* para 43 proteínas.

3.1.1.2. ENCODE

Como se ha explicado al principio de esta subsección, la inclusión de datos transcriptómicos puede ayudarnos a enfocar los esfuerzos de nuevos análisis hacia tipos de muestras en los que puede haber una probabilidad mayor de encontrar lo que se está buscando, como por ejemplo en este caso proteínas *missing*. Buscando fuentes de datos que ayudasen a completar el mapa transcriptómico a una mayor cantidad de líneas celulares y tejidos, se consideró la inclusión de datos no provenientes del proyecto *HPP*. El proyecto *ENCODE*[206] contiene una gran cantidad de experimentos transcriptómicos acerca de una gran variedad de líneas celulares. Este proyecto nació en 2003 y fue fundado por el *National Human Genome Research Institute* (*NHGRI*) con el objetivo de identificar todos los elementos funcionales del genoma humano, incluyendo elementos que actuan tanto a nivel de proteínas como de ARN, así como elementos regulatorios que controlan el funcionamiento de las células mediante la modulación de los genes. De esta forma, los grupos de investigación asociados al proyecto han realizado una gran cantidad de experimentos en este sentido, como *RNA-Seq* o *ChIP-Seq*. En este trabajo se ha analizado la expresión génica de todas las líneas celulares con datos de *RNA-Seq* procedentes de ARN de toda la célula con colas poliA+, siendo en enero de 2015 de 23 líneas celulares. Algunas de estas líneas celulares contenían más de un experimento, y en varios de estos experimentos existían varias muestras biológicas distintas, formando un total de 44 experimentos y 288 secuenciaciones realizadas en total. Todos estos datos desglosados se encuentran en la Tabla 3.1.

3.1.1.3. *Illumina Human Body Map 2.0*

En el mismo contexto de análisis proteogenómico se incluyó el proyecto *Human Body Map*, que se llevó a cabo en 2010 y consistió en la generación de perfiles transcriptómicos, con secuenciadores *Illumina HiSeq 2000*, de 16 tejidos humanos: adiposo, adrenal, del cerebro, pecho, colon, corazón, riñón, hígado, pulmón, nódulo linfático, ovario, próstata, músculo esquelético, testículos, tiroides y glóbulos blancos. Los datos de este proyecto fueron liberados en 2011, y fueron utilizados para generar modelos de genes para el humano en la versión de *Ensembl 62*. Estos datos se han publicado en *GEO* bajo la accesión *GSE30611*, en *ArrayExpress* bajo el nombre de *E-MTAB-513*, y en *SRA*[156] bajo la accesión *ERX011226*.

Tabla 3.1: Lista de las 23 líneas celulares del proyecto ENCODE utilizadas en el estudio

Línea celular	*Tejido*	*Número de experimentos incluídos*	*Número de muestras incluídas*
BJ	piel	1	2
IMR90	pulmón	1	2
K562	sangre	3	6
SK-N-SH	cerebro	1	2
NHEK	piel	3	6
HEPG2	hígado	3	6
GM12892	sangre	1	3
GM12891	sangre	1	2
HELA-S3	cervix	3	6
HSMM	músculo	2	4
H1-HESC	células madre embrionarias	5	10
GM12878	sangre	4	7
HUVEC	vaso sanguíneo	3	6
NHLF	pulmón	2	4
A549	epitelio	1	2
MONOCYTES-CD14+	monocitos	1	2
HMEC	pecho	1	1
SK-N-SH_RA	cerebro	1	2
LHCN-M2	mioblastos de músculo esquelético	2	2
CD20+	sangre	1	2
MCF-7	pecho	2	5
AG04450	pulmón	1	2
Total		**44**	**86**

3.1.1.4. *Cancer Cell Line Encyclopedia*

El proyecto *Cancer Cell Line Encyclopedia* (*CCLE*)[23] es una compilación de datos de expresión génica, variación en número de copias en cromosomas y compendio de mutaciones de 947 líneas celulares de cáncer humanas, algunas de ellas además tratadas con diferentes compuesto farmacológicos, llevado a cabo como una colaboración entre el *Broad Institute*, *Novartis Institutes for Biomedical Research* y su *Genomics Institute of the Novartis Research Foundation*. Proporciona acceso público a los perfiles de expresión génica de 917 de estas líneas celulares correspondientes a 36 tipos de cáncer diferentes utilizando la plataforma de microarrays *Affymetrix Human Genome U133 Plus 2.0*. Estos datos se encuentran publicados en *GEO* bajo la accesión *GSE36133*.

3.1.1.5. *IGC's Expression Project for Oncology*

El *International Genomics Consortium* (*IGC*) es una organización médica sin ánimo de lucro cuyo fin es el de acelerar el paso de los avances en genómica a la medicina y la industria. Este consorcio creó el proyecto *Expression Project for Oncology* (*expO*) en 2004, que consiste

en una colección de tumores, con 2158 muestras que corresponden a 156 tejidos diferentes, utilizando también la plataforma de microarrays *Affymetrix Human Genome U133 Plus 2.0*. Este conjunto de datos también se encuentra disponible en *GEO* bajo la accesión *GSE2109*.

3.1.1.6. Otros estudios

Además de todos estos datos transcriptómicos procedentes de diversos proyectos, se han incluido los datos de varias publicaciones, interesantes por su gran cantidad de muestras y diversidad de tejidos y líneas celulares. Entre estos estudios se encuentra uno que estudia la leucemia mieloide aguda[98], en el cual la técnica *RNA-Seq* fue utilizada para determinar los patrones de expresión génica en líneas celulares de este tipo de enfermedad que albergan una aberración 3q. Los datos de este estudio incluyen muestras de siete líneas celulares, las cuales están disponibles en *ArrayExpress* con el número de entrada *E-MTAB-2225*. También se han incluido los datos de la publicación de *Roth et al.*[245] que contiene los perfiles transcripcionales de varios tejidos humanos, provenientes de diez donantes fallecidos, y que fueron analizados con la plataforma de *microarrays Affymetrix Human Genome U133 Plus 2.0*. En total son 352 muestras provenientes de 65 tejidos, las cuales se encuentran almacenadas en *GEO* bajo la accesión *GSE3526*.

3.1.2. Análisis de datos transcriptómicos

El análisis de este tipo de datos permite por un lado obtener valores fiables de expresión del conjunto de genes del genoma humano, y de este modo ver cuales de las muestras podrían estar produciendo proteínas de interés. Además, algunos de estos tipos de experimentos pueden ayudarnos a comprender las características concretas de cada tipo de tejido, como mutaciones o patrones de *splicing* específicos, y poder mezclarlos con datos proteómicos para aplicar de este modo técnicas de proteogenómica.

3.1.2.1. Análisis de datos de *RNA-Seq*

Después de descargar los ficheros *FASTQ* de secuenciación de los diferentes proyectos, se siguió el flujo de análisis detallado en la Figura 3.1 para analizar este tipo de experimentos. Como primer paso, se llevó a cabo un control de calidad con la herramienta *FastQC*[32], y se eliminaron contaminantes y adaptadores residuales con la herramienta *FASTX-toolkit*[227, 147]. Las lecturas que pasaron estos controles fueron alineadas contra la versión *GRCh37* del genoma humano de referencia utilizando *TopHat*[284, 137] (versión 2.0.9). Una vez realizado el alineamiento, se utilizó la herramienta *SAMtools*[161] *rmdup* para eliminar potenciales duplicados *PCR* que puedan falsear los datos de cuantificación. Esta herramienta, utilizada específicamente con datos de secuenciación *paired-end*, en el caso de que existan múltiples parejas de lecturas que aparezcan mapeados en coordenadas idénticas, mantienen únicamente una pareja de lecturas, que será la que tenga la mayor puntuación de calidad de alineamiento, bajo la suposición de que el resto de lecturas exactamente iguales con las mismas coordenadas de mapeo serán duplicados *PCR* generados en la fase de preparación de la librería de secuenciación. En el caso de la existencia de réplicas técnicas, en este caso se unieron con la herramienta *SAMtools merge*. Posteriormente los alineamientos se ordenaron por nombre de lectura utilizando la herramienta *SAMtools sort*. El proceso de cuantificación se realizó en dos partes, una primera de conteo de fragmentos, llevada a cabo mediante el script *htseq-count* contenido en el paquete *HTSeq*[12], y utilizando anotaciones de genes de *GENCODE*[105] (versión V18), y una segunda de cálculo de lecturas normalizadas *FPKM*(fragmentos por kilobase de tránscrito por millón de fragmentos mapeados) que fue realizada mediante un script *Ruby* propio.

Los valores de lecturas normalizadas, como los *FPKM*, se suelen utilizar para poder comparar la expresión entre varios experimentos, pero es un valor no muy fiable a la hora de determinar si un gen está realmente expresado. Existen genes cuyos niveles de expresión necesarios para realizar su función son muy pequeños, pero también, a pesar de los grandes avances llevados a cabo con las tecnologías *NGS*, puede haber unos niveles de ruido apreciables en este tipo de experimentos, además de poder realizarse alineamientos incorrectos, siendo muy complicado distinguir si realmente existe expresión del gen, o si se trata de ruido. Algunos autores han propuesto lidiar con este problema sencillamente conservando un porcentaje arbitrario de los genes más expresados, pero esta estrategia no tiene en cuenta la biología del experimento, y muy probablemente filtre genes que realmente se están expresando. Es necesario el uso de métodos más robustos para analizar este tipo de datos. Existen métodos que utilizan las lecturas alineadas en zonas intergénicas para evaluar el posible ruido del *RNA-Seq*, y poder proponer unos valo-

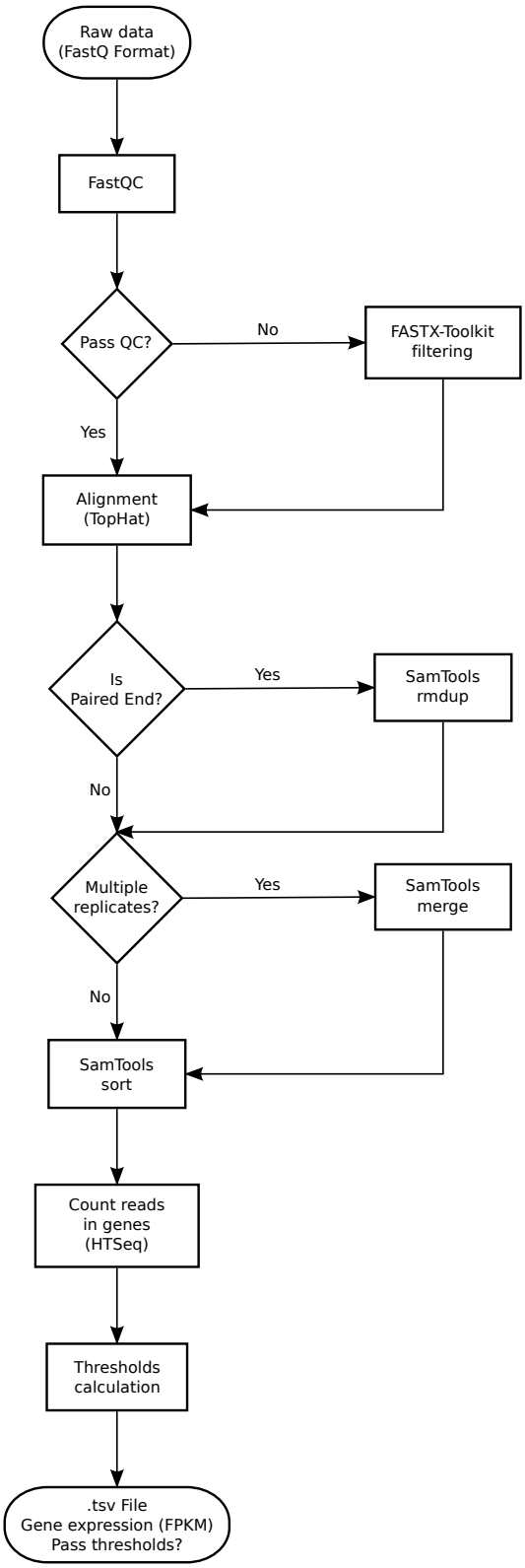


Figura 3.1: Flujo de análisis transcriptómico para *RNA-Seq*, desde los datos de partida hasta el conteo normalizado

res umbrales de expresión, tales como el explicado en el artículo de *Ramskold et al.*[234], o el método del proyecto *Bgee*[244], que está basado en el trabajo de *Hebenstreit et al.*[109].

3.1.2.1.1 Implementación de métodos de umbral para *RNA-Seq*

Con el objetivo de determinar el nivel a partir del cual consideramos un tránscrito expresado, para cada experimento *RNA-Seq* descargado y analizado, los métodos de umbrales previamente citados fueron adaptados y calculados, además de incluir gráficos con los resultados de los cálculos de umbral.

Estos dos métodos se basan en contar lecturas intergénicas e interpretarlas como ruido. En este caso se utilizaron las anotaciones de *GENCODE* (versión V18) para delimitar las zonas intergénicas, tomando los rangos existentes de los genes e invirtiéndolos para obtenerlas. En el estudio de *van Bakel et al.*[18] se afirma que la densidad de lecturas media en zonas de comienzo y fin de genes es muy alta, y que correlaciona con los genes conocidos, por lo que no deberían ser tenidas en cuenta. Como resultado de esto, se ajustaron las coordenadas de las regiones intergénicas eliminando 10 kilobases de las zonas de comienzo y fin de los genes. Finalmente se realizó un recuento de fragmentos utilizando el script previamente mencionado *htseq-count*.

En el método de *Ramskold et al.*[234], el cálculo del umbral se realiza de la siguiente manera. La información acerca de la expresión de genes y regiones intergénicas se compartimentaliza utilizando una escala logarítmica desde $log_{10}(0.01)$ hasta $log_{10}(100)$ tomando pasos de 0.1 en 0.1 en esta escala transformada, y generando cantidades acumuladas de genes y regiones intergénicas expresadas por encima de diferentes niveles. Posteriormente se calcula una tasa de falsos descubrimientos (*FDR*) para cada uno de estos niveles de expresión:

$$FDR_i = \frac{\frac{acum\_interg_i}{acum\_genes_i} * (1 - acum\_genes_i)}{1 - acum\_interg_i} \tag{3.1}$$

que es entonces utilizado para estimar el verdadero número de genes expresados en cada región a partir de los datos de expresión génica, que se calcula como la multiplicación de la densidad de genes en ese punto por la tasa de falsos descubrimientos en el mismo:

$$genes\_est_i = acum\_genes_i * FDR_i \tag{3.2}$$

A continuación, se calcula una tasa de falsos negativos (*FNR*) a partir de este número estimado de genes expresados realmente:

$$FNR_i = 1 - \frac{\sum_{j=i}^{n} genes\_est_j}{\sum_{j=1}^{n} genes\_est_j} \tag{3.3}$$

Finalmente, el umbral se calcula como la interseción entre la función generada por la *FDR* y la *FNR* en los diferentes puntos. Este método se ha implementado como un script *Ruby*, incluyendo código adicional en *R* para generar las curvas a partir de los diferentes puntos de *FDR* y *FNR* calculados, y poder conocer el punto exacto de la intersección.

El método de umbral utilizado en el proyecto de *Bgee* también utiliza la expresión de genes y zonas intergénicas para su cálculo. Los niveles de expresión se dividen en regiones, y para cada región, se calcula la proporción de genes y regiones intergénicas expresados sobre esos niveles comparado con la proporción total:

$$R_i = \frac{N\_interg\_sobre_i * N\_genes}{N\_genes\_sobre_i * N\_interg} \tag{3.4}$$

El valor de umbral se define como el mínimo valor de *i* para el cual *R* sea igual o menor que un valor fijo, que en este caso es de 0.05. Este método también se ha desarrollado como un script en *Ruby*.

3.1.2.1.2 Bases de datos para proteogenómica

Con el *dasHPPboard* también se quiere facilitar la realización de análisis proteogenómicos proporcionando bases de datos de péptidos construidas utilizando la información de variantes (*SNPs*) y nuevas uniones entre exones obtenidas de los datos de *RNA-Seq*. Estas bases de datos pueden ser utilizadas para la identificación de variantes de proteínas presentes en experimentos de proteómica de *shotgun* en la misma línea celular o tejido de los cuales se extrajeron estos *SNPs* y nuevas uniones de exones.

Se han utilizado dos aproximaciones distintas a la hora de generar las bases de datos de péptidos que contienen polimorfismos de un único aminoácido (*SAPs*) y para las de nuevas uniones entre exones. En ambos casos los datos de partida son los provenientes del análisis de *RNA-Seq*, con los datos ya filtrados y alineados.

En el caso de las bases de datos de *SAPs*, primeramente se utilizó la herramienta *SAMtools mpileup* para realizar la búsqueda de *SNPs*. El fichero binario obtenido (extensión *BCF*) fue convertido al formato *variant call format* (*VCF*) con la herramienta *BCFtools*[208] (versión 0.1.17-dev) y posteriormente filtrado con el script *vcfutils.pl* contenido dentro del paquete de *SAMtools*. Los *SNPs* con una profundidad menor que 10 lecturas o una calidad menor que 10 no se consideraron para análisis posteriores. Los *SNPs* restantes fueron anotados con la herramienta *Variant Effect Predictor*[188, 187] de *Ensembl*, que permite convertir las variantes nucleotídicas en los aminoácidos correspondientes, y además permite obtener las puntuaciones de predicción de efectos funcionales de los mismos utilizando los programas *SIFT* y *PolyPhen-2*. Esta información de predicción fue utilizada para filtrar las variantes sinónimas (sin cambio de aminoácido) utilizando scripts escritos en *R*. Posteriormente, con los variantes restantes, se extrajeron secuencias de 80 aminoácidos alrededor del péptido mutado a partir de ficheros *FASTA* de proteínas provenientes de la base de datos de *Ensembl*. La base de datos de *SAPs* incluye aquellas mutaciones que no existan en el proteoma de referencia (utilizando la versión de *Ensembl* v73). Esta aproximación proteogenómica se basa en la llevada a cabo por *Sheynkman et al.*[261].

Las bases de datos de los péptidos resultantes de nuevas uniones entre exones fueron generadas utilizando el paquete de *R customProDB*[300]. En este caso se parte de un fichero adicional generado por la herramienta *TopHat*, un fichero *BED* que contiene los sitios de unión detectados al realizar el alineamiento de secuencias. Este fichero es leído, filtrando las uniones entre exones ya conocidas. Las restantes se utilizan para obtener secuencias nucleotídicas que son traducidas en los 3 marcos de lectura diferentes dando como resultado secuencias de péptidos. Estas secuencias se comparan entonces con las de péptidos ya conocidos, filtrando los ya existentes en la referencia. En este caso, la longitud de los péptidos no es constante, depende de la longitud de los exones que se unen mediante estas nuevas uniones definidas por *TopHat*.

Estos flujos de trabajo se han implementado para utilizar los datos de *RNA-Seq* a nivel de muestra, así que en los experimentos que incluyen más de una contendrán varias bases de datos, que pueden ser agrupadas por los usuarios a su antojo, siendo necesario únicamente un paso de eliminación de redundancia de secuencias. Además se proporcionan bases de datos de referencia de secuencias contaminantes provenientes de *cRAP*[281], así como secuencias señuelo de las bases de datos de péptidos, utilizando el método de pseudoinversa con tripsina.

3.1.2.2. Análisis de datos de *microarrays*

Los experimentos de *microarrays* utilizados en este estudio fueron procesando utilizando el siguiente flujo de trabajo. Todo el flujo de trabajo se ha llevado a cabo utilizando el entorno estadístico *R/Bioconductor*[90]. La correción de ruido de fondo y la normalización de los valores del *microarray* se llevó a cabo utilizando el algoritmo *fRMA*[182] (*Frozen Robust Multichip Average*). Posteriormente a este paso de normalización, se calculó un umbral de probabilidad de expresión para cada muestra para poder distinguir las sondas expresadas de las no expresadas utilizando el algoritmo *Gene Expression Barcode*[184]. Mediante este algoritmo se estimó este valor teniendo en cuenta gran cantidad de datos de diferentes plataformas *Affymetrix* disponibles en repositorios públicos, incluyendo la plataforma *Affymetrix Human Genome U133 Plus 2.0*. El resultado de este algoritmo es un valor z calculado bajo la distribución normal de los conjuntos de sondas no expresados. Un valor de z mayor que dos fue utilizado para identificar a los genes expresados en cada muestra. Se calculó también una probabilidad de expresión del gen utilizando un clasificador bayesiano ingenuo que basa sus cálculos en la longitud de secuencia de las regiones *3'UTR*, *5'UTR* y *CDS* del gen, la probabilidad de que un gen expresado sea un gen codificante de una proteína *missing*, y el código de barras del gen para cada muestra biológica[101]. De esta manera se obtiene una puntuación con la que es posible ordenar los genes de los más probablemente expresados a los menos.

3.1.3. Análisis de datos proteómicos

3.1.3.1. Análisis de datos de proteómica de *shotgun*

A partir de ficheros de salida de los espectrómetros, y normalizados a formato *mgf* (*Mascot Generic Format*), se realizaron búsquedas contra la base de datos de proteínas humanas de *UniprotKB*[174] utilizando el motor de búsqueda *Mascot*, y posteriormente se calculó un *FDR* utilizando una base de datos señuelo, y se realizó un filtrado eliminando las identificaciones con un *FDR* menor al 1 %. A partir de estas identificaciones, se realizó una inferencia de proteínas utilizando el *software PAnalyzer*[233]. Estos experimentos se analizaron de nuevo incluyendo la posibilidad de fosforilaciones con la finalidad de buscar modificaciones postraduccionales. Se realizaron búsquedas similares pero utilizando el motor de búsqueda *X!Tandem* esta vez.

3.1.3.2. Análisis de *SRM*

Los péptidos proteotípicos fueron obtenidos de las bases de datos *PeptideAtlas*[61] y *GPMDB*[54], seleccionando las transiciones óptimas para cada péptido diana en base a los espectros *MS/MS* adquiridos previamente y a predicciones y datos de repositorios públicos. Se llevó a cabo un proceso de confirmación de los péptidos mediante el algoritmo *MIDAS*[301], y posteriormente las proteínas de este conjunto fueron detectadas utilizando al menos tres transiciones por péptido. Se pueden encontrar más detalles del procedimiento en el artículo de *Segura et al.*[256].

3.1.4. Implementación de la herramienta

El *dasHPPboard* está compuesto por la unión de diferentes componentes y tecnologías para poder clasificar, acceder y visualizar todos estos resultados de una forma modular e independiente. Antes que nada, es necesaria una estructura para poder organizar y acceder a los datos. De esta forma se ha generado una librería en la cual se han unido todos estos datos con un conjunto de metadatos extraídos de las bases de datos y proyectos de donde fueron tomados para poder organizarlos y visualizarlos adecuadamente. Los metadatos se almacenan implícitamente dentro de un árbol de directorios, en el cual cada nivel representa una característica diferente de los datos. Esta forma de almacenar los metadatos tiene la ventaja importante del ahorro en espacio de disco al aparecer de forma implícita, además de un acceso mucho más sencillo a los datos. La librería que maneja estos datos puede recorrer este árbol de directorios para realizar consultas sobre los mismos, además de poder disponer de sus metadatos asociados. También se ha diseñado un programa para poder incluir nuevos datos en el árbol de directorios de manera sencilla. Tanto la librería como el programa de inclusión de nuevos datos están escritos en *Ruby*.

El diseño del árbol de directorios utilizado para almacenar los datos funciona de forma jerárquica, como puede observarse en la Figura 3.2. Primero se dividen los datos por el proyecto al que pertenecen, que está muy ligado con el tipo de muestras que contienen, tales como tejidos o líneas celulares. En siguientes niveles de carpetas se almacenan los siguientes metadatos: la versión del genoma utilizada para analizar los datos, el nombre del tejido o línea celular y por último información acerca de los experimentos, tal como el compartimento celular analizado o el laboratorio donde se ha realizado el experimento. Tomando como base la librería de acceso a

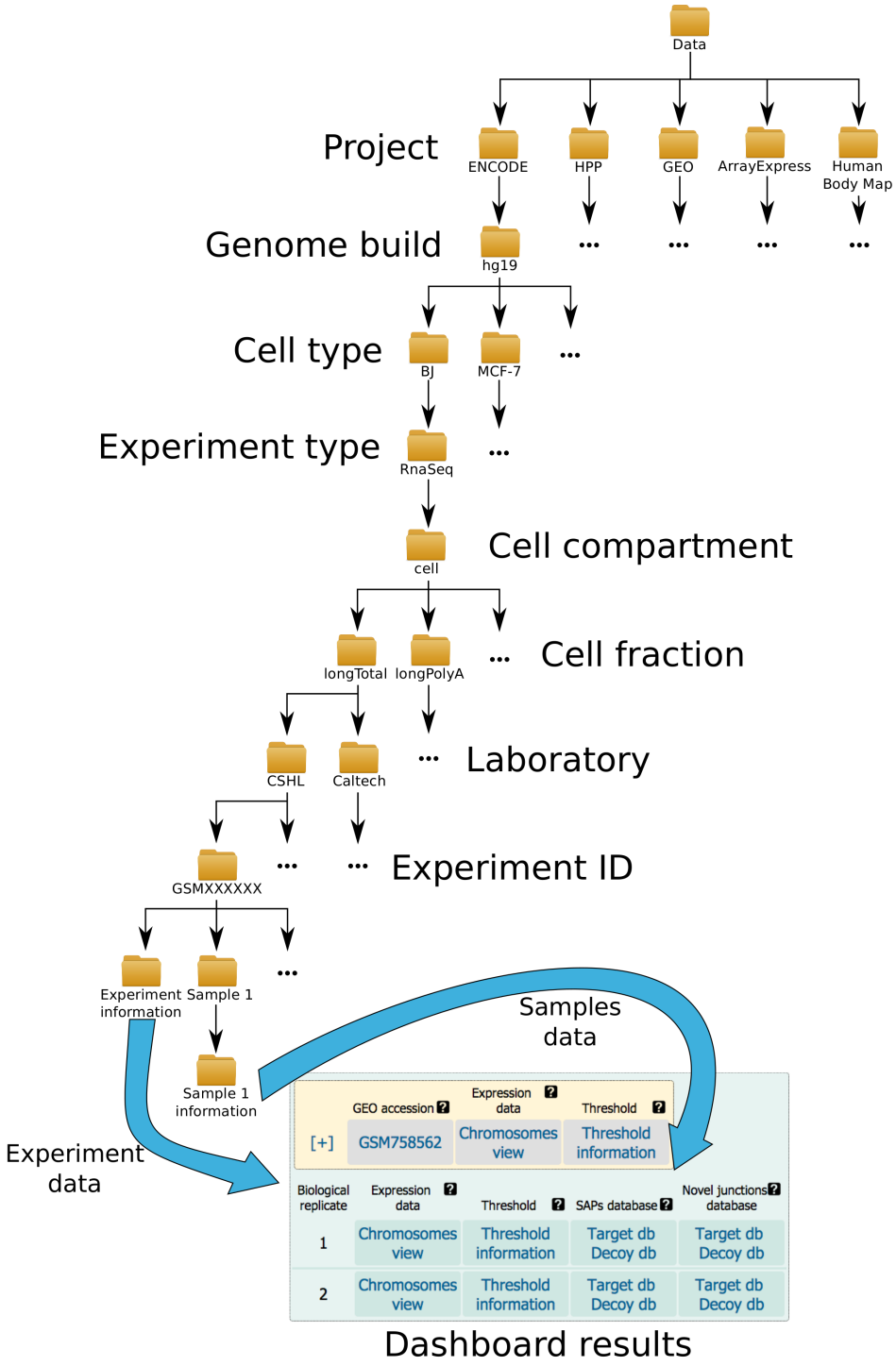


Figura 3.2: Representación gráfica de la estructura de datos interna del *dasHPPboard*

los datos mencionada previamente, y utilizando el *framework* web *Nanoc*, que genera de forma programática una visualización estática en *html* utilizando *Ruby* como base, se ha realizado el desarrollo del visor web para explorar los experimentos en el panel. Finalmente, los resultados finales de los experimentos han sido procesados y divididos por cromosomas con la finalidad de descargar la mínima cantidad de datos posible en cada consulta, siendo estos ficheros del orden de unos pocos kilobytes.

A partir de todos estos datos estáticos se ha desarrollado un módulo de búsqueda de genes y proteínas. Se ha creado una base de datos *MySQL* a partir de la información y metadatos de los experimentos, utilizando un script escrito en *Ruby*, usando el patrón de arquitectura *Active Record*[82], presente como la parte de modelo en el *framework* de desarrollo de páginas web basado en el patrón de diseño de modelo-vista-controlador[145] *Ruby on Rails*. La interfaz de búsqueda permite realizar consultas tomando como entrada identificadores de gen como los símbolos oficiales o los identificadores de *Ensembl*, e identificadores de proteína como los de *Uniprot* o los de *neXtProt*. Los resultados de las búsquedas con esta herramienta incluyen información adicional acerca del gen o proteína buscada y una tabla de experimentos, en la cual se proporcionan enlaces para su visualización en el panel del *dasHPPboard*.

3.1.5. Resultados

Hasta el momento se han generado 24GB de datos de análisis de experimentos de diversa índole, en formato texto e imágenes. Esta cantidad ingente de resultados necesitaba de una interfaz en la cual visualizarlos de una forma sencilla e intuitiva. El diseño del *dasHPPboard* nace de esta necesidad, habiéndose creado una vista en formato tabular que muestra los diferentes tipos de experimentos para los distintos tipos de tejidos o líneas celulares, mostrando la información acerca del número de experimentos y muestras para cada una de estas intersecciones. A día de hoy esta herramienta visualiza 3523 experimentos conteniendo cada uno de ellos diferentes muestras, separados en diferentes pestañas que contienen 948 líneas celulares, 77 tipos de tejidos normales y 156 tipos de tejido canceroso. Estos números se pueden encontrar más desglosados en la Tabla 3.2.

Para cada una de estas pestañas, los experimentos están organizados primero por la procedencia de los mismos, es decir, el proyecto o repositorio al que pertenecen. Posteriormente estos datos se subclasifican por el laboratorio o la serie de datos a la que pertenecen, como una forma

Tabla 3.2: Resumen del número y tipo de experimentos que se almacenan en *dasHPPboard*

			Número de líneas celulares/tejidos/tejidos cancerosos	*Número de experimentos*
Líneas celulares	**Encode**	**RnaSeq (Gingeras)**	líneas celulares = 19 tejidos = 0 tejidos cancerosos = 0 total = 14	20
		RnaSeq (Myers)	líneas celulares = 14 tejidos = 0 tejidos cancerosos = 0 total = 14	24
	HPP	**Shotgun (SpHPP)**	líneas celulares = 4 tejidos = 0 tejidos cancerosos = 0 total = 4	4
		SRM (SpHPP)	líneas celulares = 6 tejidos = 0 tejidos cancerosos = 0 total = 6	6
	ArrayExpress	**RnaSeq (E-MTAB-2225)**	líneas celulares = 7 tejidos = 0 tejidos cancerosos = 0 total = 7	7
	GEO	**GeneExpressionArray (GSE6133)**	líneas celulares = 917 tejidos = 0 tejidos cancerosos = 0 total = 917	917
Tejidos	**HBM**	**RnaSeq**	líneas celulares = 0 tejidos = 16 tejidos cancerosos = 0 total = 16	32
	GEO	**GeneExpressionArray (GSE3526)**	líneas celulares = 0 tejidos = 65 tejidos cancerosos = 0 total = 65	352
	HPP	**SRM (SpHPP)**	líneas celulares = 0 tejidos = 3 tejidos cancerosos = 0 total = 3	3
Tejidos cancerosos	**GEO**	**GeneExpressionArray (GSE2109)**	líneas celulares = 0 tejidos = 0 tejidos cancerosos = 156 total = 156	2158
Total			líneas celulares = 948 tejidos = 77 tejidos cancerosos = 156	3523

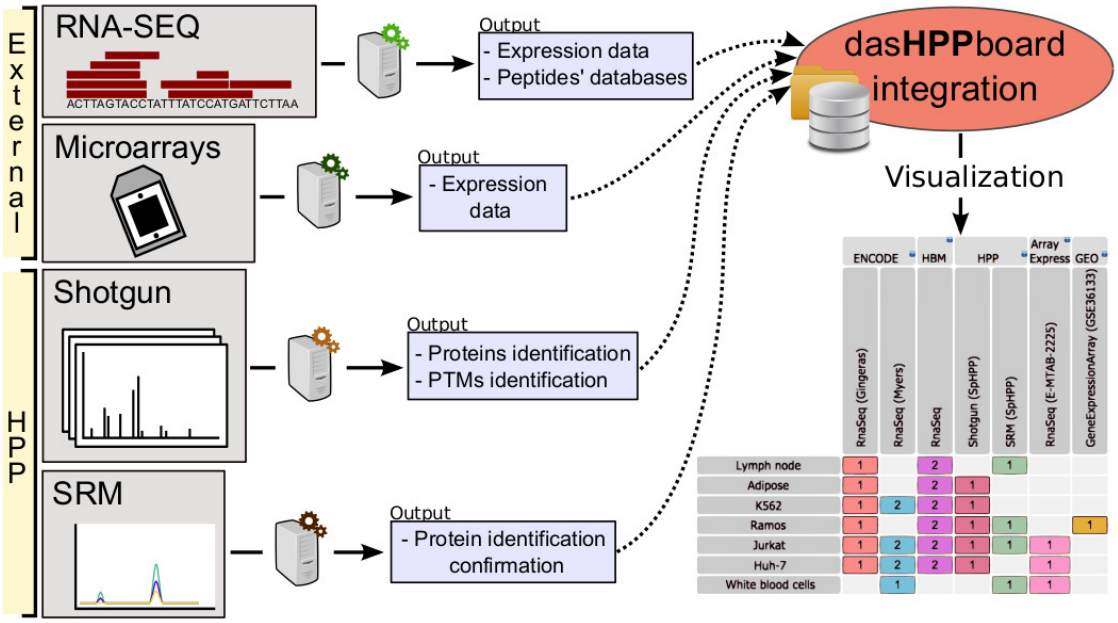


Figura 3.3: Visión global de la estructura del *dasHPPboard*, sus fuentes de datos y tipos de resultados representados

de diferenciar a los mismos dentro de un mismo proyecto y tipo de célula. Dentro de cada una de estas celdas sólo se muestra en un primer momento el número de experimentos que pertenecen a esta clasificación, habiendo añadido un código de color a estas celdas para distinguir entre diferentes tipos de experimentos. Pinchando en una de estas celdas de resultados, ésta se expande y aparece información adicional sobre los experimentos que han dado lugar a estos resultados. Esta información adicional aparece en forma de tabla y es diferente dependiendo del tipo de experimento. En la cabecera de esta nueva tabla aparece información redundante acerca del tipo de muestra y experimento, y además información acerca del compartimento y fracción celular de la muestra. Si existe información acerca del laboratorio que ha generado la muestra, también aparece aquí. Aunque existen diferentes tipos de experimentos, todos ellos han sido procesados para mostrar información a nivel de cromosoma, y en este visor se ha generado un panel de cromosomas para poder seleccionar y descargar los datos de un experimento de un determinado cromosoma. Dentro del *dasHPPboard* se pueden distinguir varios tipos de experimentos, como *RNA-Seq*, *microarrays*, proteómica de *shotgun* y *SRM*, para los cuales se han creado flujos y se han generado resultados y visualizaciones dentro del panel de visualización. Este flujo aparece representado en la Figura 3.3.

3.1.5.1. Resultados de *RNA-Seq*

Dentro del *dasHPPboard*, una vez seleccionado y expandido un resultado de este tipo, aparece una tabla en la que hay filas que contienen las diferentes muestras del experimento al que pertenecen. Para cada muestra, se visualiza la accesión de *GEO* o *ArrayExpress*, si la muestra proviene de uno de estos repositorios y el visor de cromosomas para descargar datos en forma de tabla, conteniendo las columnas descritas en la Tabla 3.3. En este tipo de experimentos se muestran también detalles del cálculo de los umbrales de expresión, incluyendo gráficos de los mismos. Como ya se comentó anteriormente, para cada experimento de *RNA-Seq* se calculan los dos métodos de umbrales explicados anteriormente. En las tablas de resultados, aparece un campo adicional haciendo referencia a estos umbrales, con valores de *HIGH QUALITY* cuando el valor de expresión génica para un gen determinado ha superado los dos valores de umbral obtenidos, *LOW QUALITY*, cuando el valor de expresión sólo ha superado uno de los dos umbrales y *NO EXPRESSION* cuando el valor de expresión no ha superado ninguno de estos dos umbrales. Aparecen también en las filas de esta tabla enlaces de descarga para las bases de péptidos generadas por los métodos proteogenómicos descritos anteriormente, conteniendo información acerca de mutaciones y nuevas uniones entre exones encontradas en la muestra. Además, si existen distintas réplicas biológicas para el experimento, aparece un enlace adicional por el cual se expande una nueva tabla y se obtiene información acerca de ellas.

3.1.5.2. Resultados de *microarrays*

Los resultados de experimentos de *microarrays* aparecen en el *dasHPPboard* de forma parecida a los resultados de *RNA-Seq*, en forma de tabla y mostrando en cada fila una muestra del experimento seleccionado. Aquí también puede aparecer la accesión *GEO* del experimento del que proviene la muestra, y un visor de cromosomas para descargar los datos, también en formato tabular, cuyos campos también aparecen descritos en la Tabla 3.3. En el caso de los *microarrays*, se incluye como valor añadido en los resultados los valores calculados de z explicados previamente, dando valores de *NA* a los genes marcados como no expresados, es decir, con un z menor que 2.

Tabla 3.3: Lista de características incluidas en las tablas de resultados de experimentos que están disponibles para su descarga en *dasHPPboard*

	RNA-Seq	***Microarrays***	***Shotgun***	***PTMs***	***SRM***
Ensembl Gene ID	X	X	X	X	X
Gene Name	X	X	X	X	X
Gene Description	X	X	X	X	X
FPKM	X				
Expression Quality	X				
neXtProt ID	X	X	X	X	X
Threshold Information	X				
Z score		X			
Protein Missing	X	X	X	X	X
Group of proteins peptide belongs to			X		
Uniprot protein isoforms			X	X	
Uniprot canonical protein ID			X	X	X
Type of protein group (unique, group of isoforms, protein family)			X		
neXtProt protein evidence			X	X	X
Protein inference category (Panalyzer)			X	X	
Mascot score			X		
Number of peptides that support the identification			X		
Peptide type				X	
Peptide sequence				X	X
PTM sequence				X	
Peptide p value				X	
Peptide q value				X	
PTM type				X	
PTM position				X	
PTM reported by Uniprot				X	
PTM reported by Phosphositeplus				X	
Retention time					X
Collision energy					X
Experiment setup					X

3.1.5.3. Resultados de proteómica de *shotgun*

En este caso, al estar extraídos los experimentos de los que provienen estos datos de la plataforma *proteomeXchange*, al expandir un resultado de este tipo aparece primeramente el código de accesión de este repositorio. También aparece la vista de cromosomas para descarga de resultados. En este caso los resultados también tienen forma tabular, e incluyen información del experimento a nivel de péptidos y proteínas, cuyas características aparecen en la Tabla 3.3. En este tipo de experimentos aparece la opción de expandirlos, mostrando los diferentes experimentos que se han consolidado para crear los resultados de *shotgun*. Para cada uno de ellos aparece información de la configuración experimental del mismo, como el instrumento y el método utilizados, además del laboratorio donde se ha realizado. También aparece una vista de cromosomas para poder descargar los *PTMs* encontrados. Estos resultados en formato tabular contienen entradas para cada combinación de proteína y *PTM*, indicando múltiples características descritas en la Tabla 3.3, incluyendo una lista que indica si cada una de las posiciones reportadas ya aparecía en *UniProt* o *PhosphoSitePlus*[117] previamente.

3.1.5.4. Resultados de *SRM*

Los resultados de experimentos de este tipo, al expandirse, contienen la vista de cromosomas para descargar los ficheros asociados al mismo. Estos ficheros de resultados, en forma de tablas, contienen diferente información asociada a la identificación de proteínas por este método que han sido listadas en la Tabla 3.3.

3.1.5.5. Resultados centrados en proteínas *missing*

De los 3523 experimentos contenidos en esta herramienta, 3510 están clasificados como transcriptómicos. El valor de esta herramienta consiste en poder mostrar información acerca de la expresión de los genes, en especial la de los genes correspondientes a proteínas *missing*, en una gran variedad de tejidos y líneas celulares. Tener de una forma centralizada toda esta información transcriptómica centrada en la expresión de genes que pueden codificar proteínas *missing* puede resultar muy interesante a la hora de buscar tejidos y líneas celulares candidatos para realizar nuevos análisis proteómicos en los que encontrarlas.

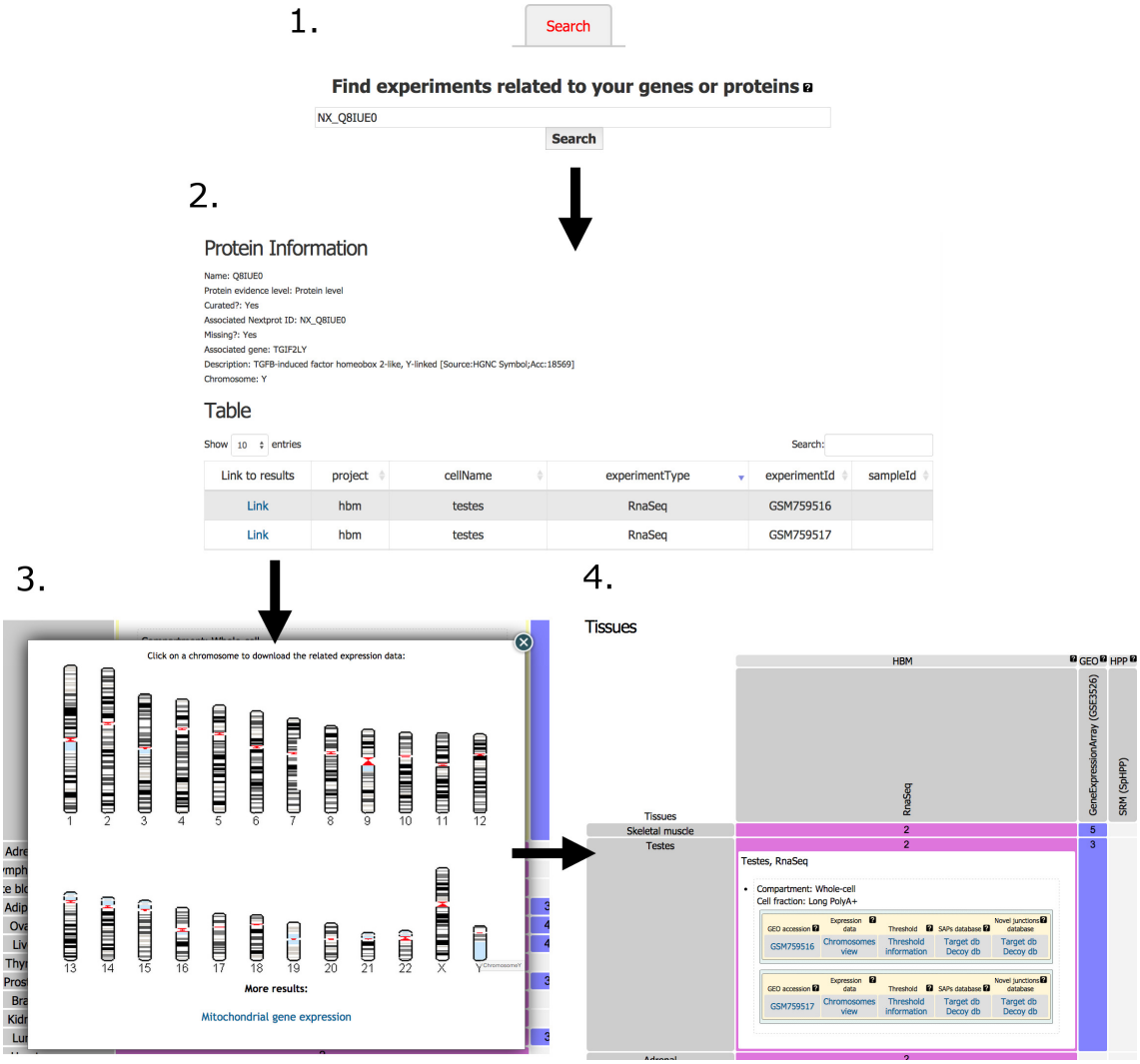


Figura 3.4: Ejemplo de uso de la sección de búsqueda del *dasHPPboard*

La sección de búsqueda del *dasHPPboard* también puede ser de gran utilidad para esta tarea, ya que realizar una búsqueda a ciegas puede ser muy laboriosa. Se puede realizar una búsqueda de cualquier proteína de interés, y comprobar si existen estudios transcriptómicos para los cuales el valor de expresión es significativo. Un ejemplo de este flujo está reflejado en la Figura 3.4. En el momento de realizar el estudio, la proteína con el identificador de *neXtProt NX_Q9IUE0* aparecía marcada como *missing* (*release* de *neXtProt* 2014-09-19). Introduciéndola en el cuadro de búsqueda, aparece una lista de resultados, entre los cuales se encuentran estudios de *RNA-Seq* para los cuales se ha encontrado nivel de expresión significativa. El tipo de muestra sobre el que han realizado estos estudios transcriptómicos es tejido testicular, el cual concuerda con la descripción presente en la base de datos *neXtProt*, indicando que podría tener un papel regulador

de otro gen que se expresa en los testículos. A partir de esta búsqueda, podemos movernos a los diferentes experimentos encontrados, y descargar las bases de datos proteogenómicas, para poder realizar estudios proteómicos mucho más dirigidos. En este caso no ha sido necesario, ya que en la siguiente versión de la base de datos *neXtProt* (de mayo de 2015) esta proteína ya ha sido marcada como codificante de proteínas basándose en evidencias proteómicas, por lo que queda demostrada la utilidad de la herramienta en este tipo de casos.

3.2. Una herramienta de enriquecimiento modular no redundante para genómica funcional (*GeneCodis3*)

Los estudios de expresión de ARN se han extendido en el área de la biología. En muchas ocasiones este tipo de estudios se llevan a cabo para tener una visión global de lo que ocurre en las células analizadas. Pero después de los análisis estadísticos de expresión diferencial, muchas veces el resultado consiste en listas de centenares de genes que pueden ser interesantes ya que su expresión ha cambiado. Debido a esta necesidad, han ido apareciendo con el tiempo un grupo de herramientas, llamadas de enriquecimiento funcional de genes, que permiten a partir de estas listas, y con información biológica adicional acerca de cada uno de estos genes por separado, obtener información biológica global acerca de la naturaleza del experimento.

La mayor parte de las herramientas de análisis de enriquecimiento de genes con anotaciones existentes están diseñadas para evaluar anotación por anotación de forma individual, perdiendo de esta forma las potenciales relaciones entre ellas. Como ya se ha explicado anteriormente en la Sección 1.1.5.1.2.1, los algoritmos de este tipo de herramientas se denominan de enriquecimiento singular. Poder encontrar relaciones entre anotaciones basándose en la co-ocurrencia puede ayudarnos a mejorar el entendimiento biológico de listas de genes asociadas a un experimento. El hecho de utilizar grupos de anotaciones para el análisis puede también añadir nuevos términos biológicos al mismo, ya que mejora la representatividad en el análisis estadístico. Como ya se ha comentado, posteriormente a las herramientas de enriquecimiento singular aparecieron entre otras, las de enriquecimiento modular, que básicamente permiten este análisis agrupando anotaciones y permiten todas las mejoras anteriormente comentadas.

La herramienta *GeneCodis*[43] apareció en 2007 con el propósito de convertirse una herramienta de referencia de análisis de enriquecimiento modular. Esta primera versión ya contenía las bases de la actual herramienta. Era una herramienta web basada en procesamiento mediante scripts *CGI Perl*. Tomaba como referencia identificadores de gen de *Entrez*, permitiendo la traducción de identificadores de algunas bases de datos de genes (símbolos de genes, identificadores de *Unigene*, etc), y como bases de datos de anotaciones utilizaba la ontología de *Gene Ontology*[16, 34], las rutas metabólicas de *KEGG*[304], motivos de secuencias de *InterPro*[80] y palabras clave de *SwissProt*[17]. Como limitaciones de esta primera implementación podríamos citar la interfaz web, basada en poder visualizar las tablas de resultados; las limitaciones basadas en el número de bases de datos utilizadas para identificadores de genes, anotaciones, y

organismos soportados; o el mayor tiempo requerido para procesar los resultados, debido a una implementación en *Perl* sobre un único servidor, lo que podría suponer no poder atender todas las peticiones de usuarios de forma simultánea. Aún así, en el año 2009, esta herramienta había realizado más de 25000 análisis, y se instaló en un servidor espejo del *Center for Bioinformatics* de la universidad de Pekín.

En este mismo año 2009, se liberó una nueva versión de la herramienta[214], totalmente renovada. Esta nueva versión mejoró la cantidad de anotaciones, añadiendo bases de datos de micro ARNs o factores de transcripción, permitiendo también realizar análisis de enriquecimiento singular además del modular previamente realizado. Se añadieron nuevos identificadores de genes para lograr una mayor compatibilidad en los análisis, y también un mayor número de organismos. El programa encargado del cálculo estadístico fue totalmente reescrito en el lenguaje *C++* para lograr una mayor eficiencia, y se añadió soporte de ejecución en un entorno *grid*, para permitir la realización de muchos más análisis de forma concurrente. Todo el *backend* de la herramienta también fue reescrito en *Ruby*, añadiendo también un acceso programático a la misma a través de servicios web *SOAP*. Por último se mejoró la interfaz web, con una más adecuada representación de los resultados, añadiendo gráficos sencillos para una mejor interpretación visual.

En esta nueva versión de la herramienta los objetivos han sido similares a los que dieron lugar a la segunda versión: incorporar la mayor cantidad de información proveniente de bases de datos para anotar los genes, y de esta manera poder realizar mejores interpretaciones de listas de genes de experimentos; mejorar la velocidad de procesamiento de los datos, para realizar análisis en el menor tiempo posible; y mejorar la apariencia de los resultados, creando diferentes formas de visualización para una mejor comprensión de los mismos.

3.2.1. Fuentes de datos para la herramienta

GeneCodis es una herramienta que toma como partida identificadores de genes y anotaciones biológicas. Existen dos grandes problemas con este tipo de información. Por un lado, existe una gran diversidad de identificadores de diferentes bases de datos para un mismo gen. Como se ha explicado anteriormente, no existe una solución óptima para este problema, por lo que se ha optado en este caso por utilizar la herramienta *BioMart*[263], ya que contiene, partiendo de la base de identificadores de genes *Ensembl*, una gran variedad de traducciones a identificadores

de genes de otras bases de datos, traducciones a identificadores de proteínas, incluso descripciones, secuencias y más características de los mismos. Internamente, se asocian todos los posibles identificadores pertenecientes a un mismo gen, y se le asigna un identificador interno, de esta forma se pretende contener los datos de asociación de genes de la forma menos sesgada posible.

Por otro lado, la información biológica en forma de anotaciones de genes, es una información muy heterogénea, distribuida a lo largo de diferentes repositorios y bases de datos, cada uno de ellos con un formato diferente. Gran parte de las herramientas de análisis de enriquecimiento de anotaciones utilizan *Gene Ontology* como base de sus análisis. Desde su primera versión *GeneCodis* viene utilizando las anotaciones de esta ontología, utilizando además una versión, denominada *GO slim*, en la cual se han eliminado términos muy específicos, para poder de esta manera dar una explicación más general, con términos que suelen contener mayor número de genes asociados. En esta versión de la herramienta se han utilizado versiones actualizadas de anotaciones ya incluidas en la segunda versión. Por ejemplo, las rutas metabólicas de *KEGG*[304], información estructural de dominios de proteínas procedente de *InterPro*[80], información regulatoria de dianas de micro ARNs procedente de *mirBase*[143], o información de factores de transcripción procedente de *TRANSFAC*[181] almacenada en *MsigDB*[273] y *YEASTRACT*[1] en el caso de la levadura. Como novedad, en esta nueva versión se han añadido enfermedades relacionadas con genes provenientes de *OMIM*[10], información acerca del efecto en la actividad de los genes bajo el efecto de diferentes fármacos proveniente de *PharmGKB*[111, 8], rutas metabólicas procedentes de *Panther*[197] e identificadores de artículos de *Pubmed* asociadas a genes. Internamente, después de manipular los ficheros de las diferentes bases de datos, se obtiene una tabla de asociación entre términos y genes, la cual es procesada para que los identificadores de genes de estas tablas sean los códigos internos generados en el paso anterior.

3.2.2. Algoritmo para enriquecimiento

GeneCodis realiza dos tipos de análisis de enriquecimiento distintos, modular y singular. El análisis modular permite asociar términos entre sí para realizar un análisis con un mejor sentido biológico. Para poder asociar términos se ha utilizado el algoritmo *apriori*[4], que permite encontrar de forma eficiente conjuntos de elementos frecuentes para generar reglas de asociación entre ellos. Este algoritmo se diseñó para identificar conjuntos de elementos que son subconjuntos de al menos un número mínimo de transacciones en una base de datos. En este caso el

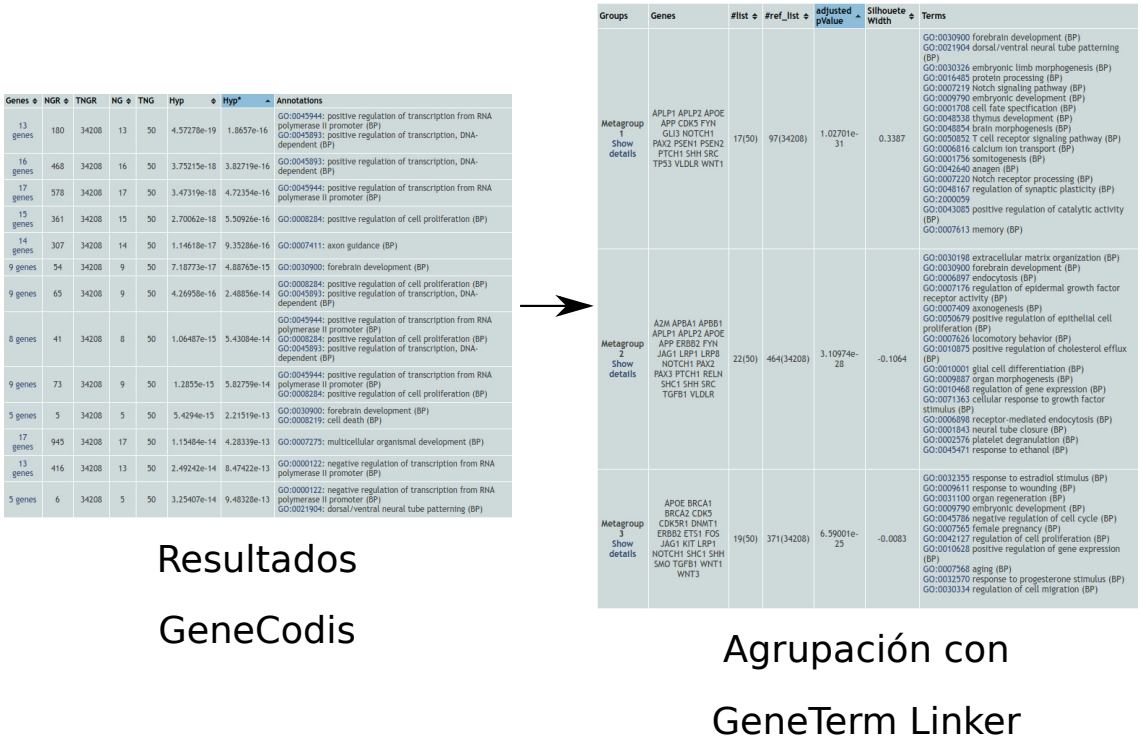
procedimiento determina conjuntos de términos que comparten al menos un número mínimo de genes, comenzando por encontrar términos únicos, utilizando esa salida para generar conjuntos de términos cada vez más grandes que compartan esos genes, hasta que no se puedan aumentar más esos conjuntos.

Posteriormente, una vez generados estos conjuntos, el análisis estadístico de enriquecimiento es equivalente. Se cuentan las ocurrencias de cada grupo de anotaciones en la lista de genes proporcionada como entrada y en la lista de genes de referencia. Por defecto *GeneCodis3* usa como lista de referencia todos los genes anotados con términos para las categorías de términos seleccionadas para ese organismo en concreto, utilizando como base el conjunto de genes de *Ensembl*. Entonces se realiza un test para encontrar las categorías enriquecidas, el cual puede ser una distribución hipergeométrica o una prueba de independencia χ^2. Posteriormente, los p-valores generados por uno de estos tests son ajustados por contraste múltiple pudiendo elegir entre dos métodos diferentes. El primero de ellos es una corrección basada en simulación, en la que seleccionan aleatoriamente subconjuntos de genes de la lista de referencia del tamaño de la lista de entrada, y se calculan p-valores para cada conjunto de categorías enriquecidas, realizando este proceso 10000 veces, y corrigiendo entonces el p-valor original calculando la fracción de simulaciones en las el p-valor para esa categoría que es mejor que el original. El segundo método sería el de falsos descubrimientos propuesto por *Benjamini* y *Hochberg*[27].

3.2.3. Refinado de resultados

En muchas ocasiones, a pesar de la utilidad de los análisis de enriquecimiento, se generan listas muy largas que pueden confundir a la hora de realizar la interpretación biológica. Muchas veces esto ocurre por la redundancia de los términos biológicos que aparecen en gran cantidad de repositorios. También existen términos que pueden ser demasiado genéricos como para proporcionar información útil.

En esta versión de la herramienta se ha incluido un reciente método denominado *GeneTerm Linker*[81] cuya finalidad es intentar agrupar los resultados de enriquecimiento buscando módulos de genes y términos coherentes entre sí. Este método realiza un filtrado y agrupación de términos y genes para después hacer un refinamiento y así eliminar redundancias en los grupos generados, generando al final un conjunto de genes y anotaciones agrupados en metagrupos. Finalmente se realizan una serie de tests estadísticos para comprobar la significancia y cohe-



Genes	NGR	TNGR	NG	TNG	Hyp	Hyp*	Annotations
13 genes	180	34208	13	50	4.57278e-19	1.8657e-16	GO:0045944: positive regulation of transcription from RNA polymerase II promoter (BP) GO:0045893: positive regulation of transcription, DNA-dependent (BP)
16 genes	468	34208	16	50	3.75215e-18	3.82719e-16	GO:0045893: positive regulation of transcription, DNA-dependent (BP)
17 genes	578	34208	17	50	3.47319e-18	4.72354e-16	GO:0045944: positive regulation of transcription from RNA polymerase II promoter (BP)
15 genes	361	34208	15	50	2.70062e-18	5.50926e-16	GO:0008284: positive regulation of cell proliferation (BP)
14 genes	307	34208	14	50	1.14618e-17	9.35286e-16	GO:0007411: axon guidance (BP)
9 genes	54	34208	9	50	7.18773e-17	4.88765e-15	GO:0030900: forebrain development (BP)
9 genes	65	34208	9	50	4.26958e-16	2.48856e-14	GO:0008284: positive regulation of cell proliferation (BP) GO:0045893: positive regulation of transcription, DNA-dependent (BP)
8 genes	41	34208	8	50	1.06487e-15	5.43084e-14	GO:0045944: positive regulation of transcription from RNA polymerase II promoter (BP) GO:0008284: positive regulation of cell proliferation (BP) GO:0045893: positive regulation of transcription, DNA-dependent (BP)
9 genes	73	34208	9	50	1.2855e-15	5.82759e-14	GO:0045944: positive regulation of transcription from RNA polymerase II promoter (BP) GO:0008284: positive regulation of cell proliferation (BP)
5 genes	5	34208	5	50	5.4294e-15	2.21519e-13	GO:0030900: forebrain development (BP) GO:0008219: cell death (BP)
17 genes	945	34208	17	50	1.15484e-14	4.28339e-13	GO:0007275: multicellular organismal development (BP)
13 genes	416	34208	13	50	2.49242e-14	8.47422e-13	GO:0000122: negative regulation of transcription from RNA polymerase II promoter (BP)
5 genes	6	34208	5	50	3.25407e-14	9.48328e-13	GO:0000122: negative regulation of transcription from RNA polymerase II promoter (BP) GO:0021904: dorsal/ventral neural tube patterning (BP)

Groups	Genes	#list	#ref_list	adjusted pValue	Silhouete Width	Terms
Metagroup 1 Show details	APLP1 APLP2 APOE APP CDK5 FYN GLI3 NOTCH1 PAX2 PSEN1 PSEN2 PTCH1 SHH SRC TP53 VLDLR WNT1	17(50)	97(34208)	1.02701e-31	0.3387	GO:0030900 forebrain development (BP) GO:0021904 dorsal/ventral neural tube patterning (BP) GO:0030326 embryonic limb morphogenesis (BP) GO:0016485 protein processing (BP) GO:0007219 Notch signaling pathway (BP) GO:0009790 embryonic development (BP) GO:0001708 cell fate specification (BP) GO:0048538 thymus development (BP) GO:0048854 brain morphogenesis (BP) GO:0050852 T cell receptor signaling pathway (BP) GO:0006816 calcium ion transport (BP) GO:0001756 somitogenesis (BP) GO:0042640 anagen (BP) GO:0007220 Notch receptor processing (BP) GO:0048167 regulation of synaptic plasticity (BP) GO:2000059 GO:0043085 positive regulation of catalytic activity (BP) GO:0007613 memory (BP)
Metagroup 2 Show details	A2M APBA1 APBB1 APLP1 APLP2 APOE APP ERBB2 FYN JAG1 LRP1 LRP8 NOTCH1 PAX2 PAX3 PTCH1 RELN SHC1 SHH SRC TGFB1 VLDLR	22(50)	464(34208)	3.10974e-28	-0.1064	GO:0030198 extracellular matrix organization (BP) GO:0030900 forebrain development (BP) GO:0006897 endocytosis (BP) GO:0007176 regulation of epidermal growth factor receptor activity (BP) GO:0007409 axonogenesis (BP) GO:0050679 positive regulation of epithelial cell proliferation (BP) GO:0007626 locomotory behavior (BP) GO:0010875 positive regulation of cholesterol efflux (BP) GO:0010001 glial cell differentiation (BP) GO:0009887 organ morphogenesis (BP) GO:0010468 regulation of gene expression (BP) GO:0071363 cellular response to growth factor stimulus (BP) GO:0006898 receptor-mediated endocytosis (BP) GO:0001843 neural tube closure (BP) GO:0002576 platelet degranulation (BP) GO:0045471 response to ethanol (BP)
Metagroup 3 Show details	APOE BRCA1 BRCA2 CDK5 CDK5R1 DNMT1 ERBB2 ETS1 FOS JAG1 KIT LRP1 NOTCH1 SHC1 SHH SMO TGFB1 WNT1 WNT3	19(50)	371(34208)	6.59001e-25	-0.0083	GO:0032355 response to estradiol stimulus (BP) GO:0009611 response to wounding (BP) GO:0031100 organ regeneration (BP) GO:0009790 embryonic development (BP) GO:0045786 negative regulation of cell cycle (BP) GO:0007565 female pregnancy (BP) GO:0042127 regulation of cell proliferation (BP) GO:0010628 positive regulation of gene expression (BP) GO:0007568 aging (BP) GO:0032570 response to progesterone stimulus (BP) GO:0030334 regulation of cell migration (BP)

Figura 3.5: Ejemplo de refinado de resultados mediante el método *GeneTerm Linker*.

rencia de estos metagrupos. Esta opción aparece en los resultados generados, ya que requiere un paso adicional de cómputo. El resultado, visible en el ejemplo de la Figura 3.5 consiste en una tabla en la que aparecen los metagrupos junto con su significancia estadística, pudiendo ver detalles adicionales de coherencia de los mismos, y los resultados de *GeneCodis* a partir de los cuales se han generado.

3.2.4. Análisis comparativo

La mayor parte de herramientas de análisis de enriquecimiento trabajan únicamente con una lista de genes. Pero en muchos estudios se comparan diferentes condiciones obteniendo varias listas. En estas ocasiones, es importante encontrar las diferencias o similitudes funcionales entre estas listas. Tradicionalmente se han realizado análisis por separado de las listas, y comparado de forma manual posteriormente los resultados.

En esta versión se ha desarrollado un análisis comparativo que permite el análisis de dos listas de genes diferentes al mismo tiempo. Se realizan análisis de enriquecimiento modular y

singular para todas las anotaciones marcadas, para cinco listas de genes en total, que serían las dos listas originales, la lista formada por la intersección de estas dos, y los genes exclusivos de cada lista. Estas operaciones se realizan en el espacio de los genes y no de las anotaciones para garantizar la significancia estadística de los resultados.

La forma de de mostrar los resultados es similar a la del análisis de una sola lista, con la adición de un diagrama interactivo de *Venn* para cada categoría de anotaciones que permite seleccionar el conjunto de genes del que se quieren visualizar los resultados.

3.2.5. Caso de uso

Para mostrar la utilidad de *GeneCodis*, se han analizado datos de genes diferencialmente expresados en la línea celular *WI-38* correspondiente a fibroblastos de pulmón humano expuestos al compuesto carcinógeno α-benzopireno [65]. En este estudio se expusieron cultivos de esta línea celular a tres distintas concentraciones de α-benzopireno (0.1, 0.5 y 1μM), y posteriormente se midieron diferencias de expresión génica utilizando *microarrays*, obteniendo 384, 972 y 837 genes diferencialmente expresados. Un esquema de los resultados de este análisis se muestra en la Figura 3.6.

En este caso se ha realizado un análisis comparativo, utilizando como listas los genes sobreexpresados e inhibidos significativamente para las tres concentraciones del compuesto. Entre otros, se ha realizado un análisis utilizando como base de datos de anotaciones los procesos biológicos de *Gene Ontology*. Como se puede observar, los genes inhibidos aparecen relacionados sobre todo con el ciclo celular, de lo que se deduce que probablemente el ciclo celular en los cultivos expuestos a α-benzopireno detienen los procesos asociados a división celular. También aparecen inhibidos genes relacionados con reparación del ADN. Los genes sobreexpresados aparecen principalmente asociados a procesos de respuesta a estrés, aceleración del metabolismo y angiogénesis. Todos estos datos pueden apreciarse a primera vista en las nubes de términos, o de una forma más exhaustiva en las tablas de resultados. Por último, se ha aplicado *GeneTerm Linker* a estos resultados de enriquecimiento, agrupando anotaciones y genes en metagrupos con una mayor coherencia biológica, pudiendo explicar así de una forma mucho más sencilla los diferentes procesos asociados a nuestros genes.

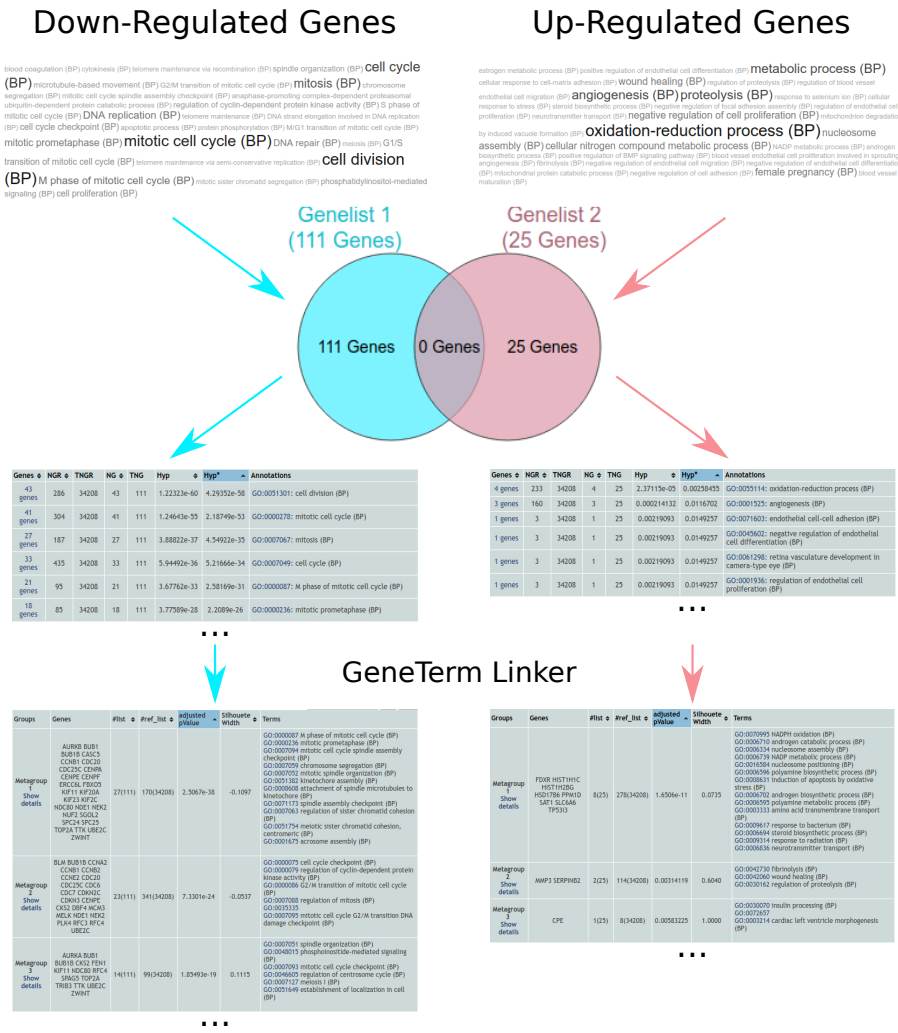


Figura 3.6: Comparación de genes diferencialmente regulados y su análisis funcional con *GeneCodis*.

3.2.6. Implementación

GeneCodis está compuesto de varias capas de *software*. La capa más interna está compuesta por código *C++* que realiza los tests estadísticos, además de generar las combinaciones de términos biológicos para el análisis modular[36]. Por encima de esta capa, aparece una librería escrita en *Ruby* para encapsular el código *C++*, donde aparecen funciones para el acceso a los genes, sus traducciones y a los términos biológicos. Esta librería es la encargada de generar también la entrada apropiada para lanzar el código *C++*, que además puede ejecutarse de forma local, o en un *cluster* de computadoras, para poder así liberar al servidor web de carga computacional. La librería también hace uso de la ejecución multihilo para minimizar el tiempo

de generación y lanzamiento de datos. Esta librería se ha implementado para poder ser llamada como un servicio web, para unificar de esta forma su utilización vía web o programática. Por último, la capa web de la aplicación se ha desarrollado con *Ruby on Rails*, uno de los *framework* web más ampliamente utilizados en la actualidad.

3.2.7. Resultados

Desde el lanzamiento de esta tercera versión de *GeneCodis*, la herramienta ha sido utilizada en más de 67000 ocasiones por unos 19000 usuarios distintos, y con una duración de visita de más de 13 minutos por sesión, lo cual refleja el nivel de popularidad que ha alcanzado esta herramienta. Aparte de las novedades metodológicas introducidas en esta versión, mencionadas previamente, se han introducido mejoras, sobre todo en la parte gráfica, para proporcionar a los usuarios datos de forma más intuitiva. Primero de todo, se ha simplificado el formulario de entrada, dejando visibles los elementos imprescindibles para poder ejecutar un análisis, y simplificando y ocultando las opciones avanzadas, tales como la selección de la lista de genes de referencia, o los tests estadísticos utilizados para calcular los p-valores y sus correcciones.

La visualización de los resultados también se ha mejorado, añadiendo a las tablas de resultados de enriquecimiento mayor interactividad, permitiendo su ordenación y filtrado. Para esta versión además se han añadido nuevos elementos visuales, como una nube de términos por tabla de resultados, que contiene los treinta términos más significativos de los resultados. Los tamaños de los términos varían dependiendo del número de genes asociados a cada uno de ellos. De esta forma se puede realizar una muy rápida interpretación de la parte más relevante de los resultados del análisis. Los gráficos añadidos en la segunda versión de la herramienta también se han renovado, consistiendo en una gráfica en forma de tarta y otra de columnas, pudiendo ser modificadas en tiempo real mediante filtros. Esta mayor interactividad en la visualización de los resultados, junto con la adición del algoritmo de *GeneTerm Linker* permite una mejor interpretación de los análisis. La incorporación de los análisis comparativos también se ha diseñado en este sentido, intentando actualizar la herramienta para un uso más cómodo con resultados de experimentos de alto rendimiento. Todas estas mejoras aparecen esquematizadas en la Figura 3.7.

Esta nueva funcionalidad, junto con la mejora en la implementación, haciendo uso de la programación con hilos para minimizar los tiempos de ejecución en partes del código que no

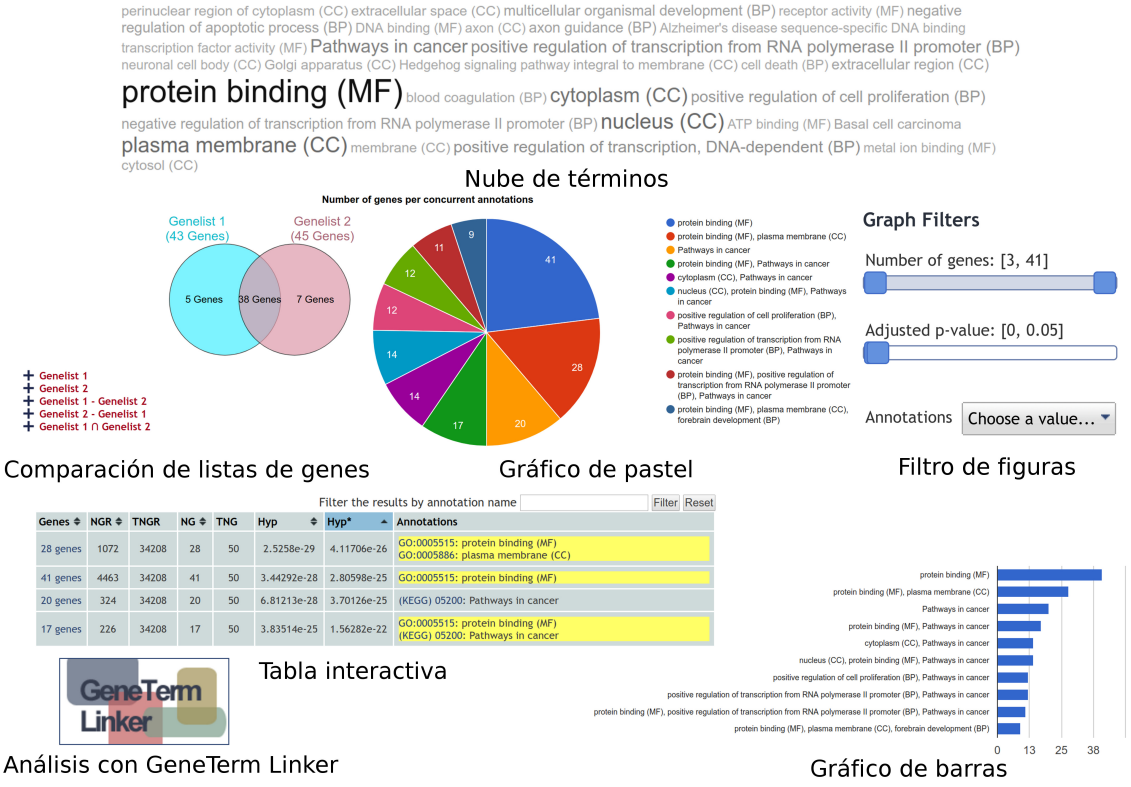


Figura 3.7: Mejoras en la visualización y análisis de datos integradas en esta versión de *GeneCodis*.

habían sido paralelizadas en anteriores versiones de la herramienta, y optimizando la configuración y lanzamiento de los trabajos en cluster para maximizar la ejecución de varios trabajos en paralelo, son las principales características que definen a esta tercera versión de la herramienta.

3.3. Mejorando las predicciones en interacciones entre ARN mensajeros y micro ARNs (*m^{3}RNA*)

A pesar de la mejora en las tecnologías de secuenciación masiva, que han permitido la puesta en marcha de métodos para descubrir nuevas interacciones entre mensajeros y micro ARNs, se presupone que existen aún muchas por descubrir, ya que los micro ARN están relacionados con la regulación de prácticamente todos los genes codificantes de proteínas[83]. Gran cantidad de grupos de investigación han dedicado esfuerzos a estudiar la naturaleza de estas interacciones para intentar comprender las características que son más importantes para que se produzcan, y así poder crear modelos que las tengan en cuenta y poder realizar predicciones que intenten finalmente explicar cómo afectan estas uniones a la expresión génica. Ya se ha explicado en la Sección 1.1.5.3.1 las características que se tienen en cuenta a la hora de realizar predicciones, y los diferentes métodos utilizados.

La existencia de un gran número de algoritmos y bases de datos de predicciones dificulta a los investigadores el proceso de buscar si existen genes diana para un determinado micro ARN, y si existen, cual es más probable que realmente se una, debido a la heterogeneidad de formatos de ficheros y diferencias en los identificadores, tanto en los genes como en los micro ARNs. El experimentalista tendría o bien que elegir una base de datos sin tener muy claro cual es la mejor, o tendría que intentar recopilar datos de diferentes bases de datos, con el tiempo que eso conlleva. La dificultad adicional existente con los identificadores puede hacer que perdamos posibles interacciones interesantes. Existen varios tipos de identificadores de micro ARNs, los cuales pueden ser muy confusos de utilizar. Normalmente siguen la convención de poner primero el código en tres letras del organismo, seguido de tres letras que pueden ser *miR* o *mir* dependiendo de si se trata de un micro ARN maduro o precursor, y posteriormente se añade un número, que se proporciona secuencialmente según su descubrimiento. Se añaden sufijos en el caso de que un micro ARN maduro pueda provenir de diferentes secuencias precursoras, y además existen diferentes nomenclaturas para denominar a los productos maduros predominantes u opuestos, nomenclaturas que además han ido cambiando con el tiempo. La base de datos *miRBase*[11] ha unificado estas diferentes nomenclaturas proporcionando un nombre único para cada secuencia.

Ante esta situación se propone el siguiente trabajo, consistente en una metodología que combina y proporciona nuevas puntuaciones a las interacciones provenientes de diferentes bases de datos. De esta forma se proporciona a la comunidad un recurso centralizado que intenta aunar

las mejores características de cada uno de los recursos existentes previamente.

3.3.1. Recursos utilizados

Como ya se ha mencionado, en este trabajo se ha intentado aunar gran cantidad de información de diferentes bases de datos de interacciones, tanto experimentalmente validadas como provenientes de predicciones. Como bases de datos experimentales, se han utilizado las previamente mencionadas: *miRWalk*, *miRecords*, *TarBase* y *miRTarBase*. Todas las interacciones presentes en estas bases de datos se han unido para conformar el estándar contra el que comparar las interacciones de las bases de datos de predicción.

Las bases de datos predictivas incluidas son: *EIMMo*[85], *DIANA-microT*[175], *Microcosm*[97], *microRNA.org*[31], *TargetScan*[158], *PITA*[134], *miRWalk-predictive*[68] y *TargetSpy*[272]. *EIMMo* busca posibles sitios de unión para un micro ARN en genes de cuatro especies distintas, y dependiendo del número de especies en los que haya habido algún posible sitio de unión, utiliza estadística bayesiana para calcular la probabilidad de conservación de la semilla. En el caso de *DIANA-microT*, primero se busca complementariedad encontrando semillas de 7 a 9 nucleótidos, o de 6 con la posibilidad de un *wobble*, en la región de la *UTR* del gen, otorgando entonces una puntuación a cada sitio comparándola con un conjunto de interacciones ya identificadas, para finalmente ponderar estas puntuaciones. Tanto *Microcosm* como *microRNA.org* utilizan el algoritmo *miRanda* explicado anteriormente. *TargetScan* se basa en el algoritmo explicado en la Sección 1.1.5.3.1.1. *PITA* escanea las *UTR* de los genes y puntúa cada sitio utilizando el método de *Kertesz et al.*[134]. El algoritmo en que se basa *miRWalk-predictive* busca la complementariedad más larga entre los micro ARNs y las secuencias de genes buscando primero en la semilla. Los sitios encontrados se clasifican por la región del gen donde se ha producido el alineamiento, y por último se calcula la distribución de probabilidad de la ocurrencia de emparejamientos aleatorios de una subsecuencia mediante una distribución de *Poisson*, asociando las mejores interacciones con probabilidades más bajas. Finalmente, *TargetSpy* utiliza como entrada *3'UTRs* y secuencias de micro ARNs, se buscan sitios complementarios, y los puntúa mediante un clasificador que estima la calidad de las características mediante el método *ReliefF*[142].

3.3.1.1. Normalización de bases de datos

Existe una gran heterogeneidad en cuanto a identificadores de genes y micro ARNs, de hecho existen incluso bases de datos que disponen de identificadores de tránscritos, una forma mucho más específica de definir la interacción. En este trabajo se ha optado por utilizar identificadores de *miRBase* y *Ensembl* como identificadores comunes, para poder comparar e integrar bases de datos. Para traducir del formato de genes original de la base de datos a *Ensembl*, se ha utilizado la herramienta *BioMart*[263]. En el caso de los identificadores de micro ARNs, *miRBase* dispone en su sección de descargas de un fichero de diccionario para convertir del resto de notaciones al formato de identificador de *miRBase*. Finalmente, para poder comparar de un mejor modo estas bases de datos, se han normalizado las puntuaciones de cada base de datos predictiva, escalando todas las interacciones de 0 a 1, teniendo una puntuación más cercana a 1 la interacción con mayor probabilidad de ser validada experimentalmente. Se probaron otras estrategias de normalización, como la resultante de sustituir las puntuaciones por uno menos la función de densidad acumulada evaluada hasta el punto en el que se encuentra cada una de las puntuaciones, intentando de esta manera eliminar el efecto de la existencia de zonas en las que un pequeño rango de puntuaciones concentran gran densidad de interacciones. Esta puntuación dependiente del rango de la interacción finalmente no fue utilizada, ya que en las pruebas realizadas no mostró ninguna mejora con el método de escalado.

3.3.2. Rendimiento de bases de datos predictivas

Antes de comenzar a realizar ningún tipo de combinación, un primer paso fue comprobar la fiabilidad de las bases de datos predictivas, de una forma muy similar a la descrita en [205]. Se utilizó todo el conjunto de interacciones experimentalmente validadas para medir el enriquecimiento en éstas dentro de cada una de las bases de datos predictivas mediante una distribución hipergeométrica. Para cada base de datos predictiva, se ordenaron las interacciones por su valor de puntuación, y se hizo un test hipergeométrico para valores diferentes de umbral en las puntuaciones. Se determinó el conjunto de interacciones más enriquecido eligiendo el umbral asociado a un p-valor más bajo del test. Este p-valor es una medida de enriquecimiento en interacciones experimentalmente validadas en estas bases de datos. Al obtener p-valores muy bajos, era muy probable cometer errores de redondeo, por lo que se utilizó la aproximación propuesta en [164]. Los resultados están incluidos en la Tabla 3.4.

Tabla 3.4: Fiabilidad de las diferentes bases de datos de predicciones de interacciones, junto con los nuevos métodos propuestos

Método	*puntuación z*	*Nº interacciones mínimo z*	*Nº interacciones BBDD*	*Nº interacciones validadas*	*Proporción interacciones validadas*	*% interacciones para mínimo z*
WSP	-84.52	123589	4669137	4286	9.18E-04	6.94
LRS	-89.27	163829	4669137	4286	8.18E-04	9.2
EiMMo	-61.87	191582	1781671	2949	1.66E-03	10.75
DIANA-microT	-54.51	269525	2889574	3010	1.31E-03	11.77
microrna.org	-21.2	134227	737379	2685	3.64E-03	18.2
microcosm	-17.99	6035	352016	784	2.23E-03	1.71
PITA	-15.2	75683	206722	1425	6.89E-03	36.61
TargetSpy	-14	178114	300000	653	2.18E-03	59.37
miRWalk	-9.92	422089	780000	1243	1.59E-03	54.11
TargetScan	-9.29	19491	132809	1832	1.38E-02	14.68
mirTarget	-5.08	149088	691265	234	3.39E-04	21.57

Estos datos mostrados pueden darnos una idea de lo bien que se comportan estas bases de datos a la hora de predecir, pero necesitamos también alguna forma de comparación entre ellas. En el área del aprendizaje máquina, el área bajo una curva *ROC* (*Receiver Operating Characteristic*) es una de las aproximaciones más usadas para medir y comparar rendimientos. En una curva *ROC* se comparan la razón de verdaderos positivos (*VPR*) o sensibilidad y la razón de falsos positivos (*FPR*), que sería equivalente a uno menos la especificidad. Estas razones se calculan del siguiente modo:

$$VPR = \frac{VP}{VP + FN} \tag{3.5}$$

$$FPR = \frac{FP}{FP + VN} \tag{3.6}$$

Donde *VP* serían los verdaderos positivos, *FN* los falsos negativos, *FP* los falsos positivos y *VN* los verdaderos negativos. En el caso de predicción de micro ARNs, estos parámetros se calculan como sigue. Una interacción se considera: *VP* en caso de que se haya predicho y esté validada; *VN* en caso de que la interacción ni se haya predicho ni esté dentro de las bases de datos experimentales; *FP* en caso de que la interacción se haya predicho pero no se encuentre validada; y *FN* en caso de que la interacción se haya determinado experimentalmente pero no se encuentre entre las predicciones. Cada punto de la curva *ROC* se obtiene utilizando diferentes valores umbral para las puntuaciones normalizadas. La Figura 3.8 muestra curvas *ROC* para todos los algoritmos predictivos.

Esta forma de medir el rendimiento no es perfecta, ya que no todas las interacciones existentes han podido ser validadas, y hay muy pocos conjuntos de interacciones que se han probado y que no se hayan validado, para servir de referencia a los verdaderos negativos. De esta forma este conjunto de *VN*, y parte de los *FP* no han podido ser correctamente clasificados. Por lo tanto realizar comparaciones con valores mal clasificados puede dar lugar a resultados erróneos. Una aproximación alternativa puede ser la de utilizar las curvas de precisión y sensibilidad. Aquí, la

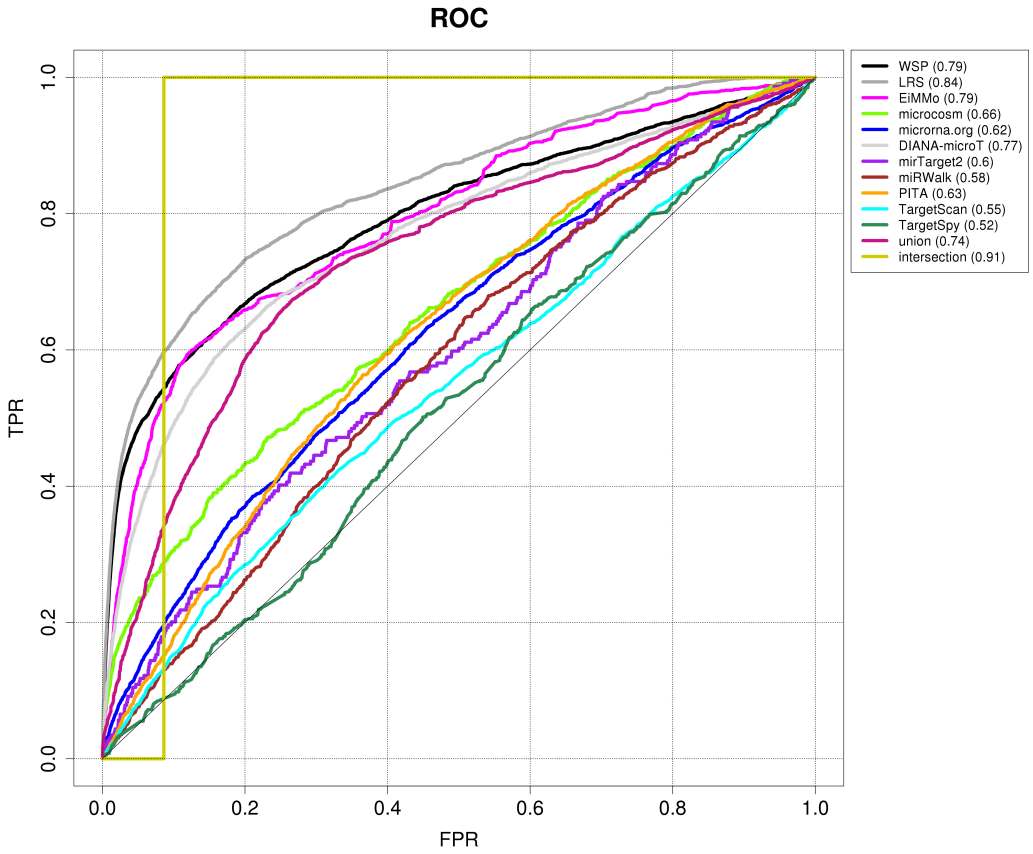


Figura 3.8: Curvas *ROC* para métodos predictivos incluidos en este trabajo junto con los nuevos algoritmos propuestos.

precisión, que se calcula como sigue:

$$Precision = \frac{VP}{VP + FP} \tag{3.7}$$

se compara con la razón de verdaderos positivos o sensibilidad. Existe una relación entre las curvas *ROC* y las curvas de precisión y sensibilidad [57], por lo que las limitaciones que se han expuesto para las curvas *ROC* también afectan a este tipo de medidas. Para poder comparar de un mejor modo, y así complementar la información procedente de las curvas *ROC*, se introduce en este trabajo el concepto de curva de precisión. Para cada base de datos predictiva, las interacciones se ordenan pos sus puntuaciones normalizadas en orden descendiente, y se determinan los valores de precisión acumulada en cada punto. La gráfica se construye utilizando el rango de las interacciones como eje horizontal, y la precisión acumulada hasta ese punto en el eje vertical. La curva resultante, generada en la Figura 3.9 para las mismas bases de datos que en el caso de la curva *ROC* anterior, muestra la proporción de las interacciones predichas que han

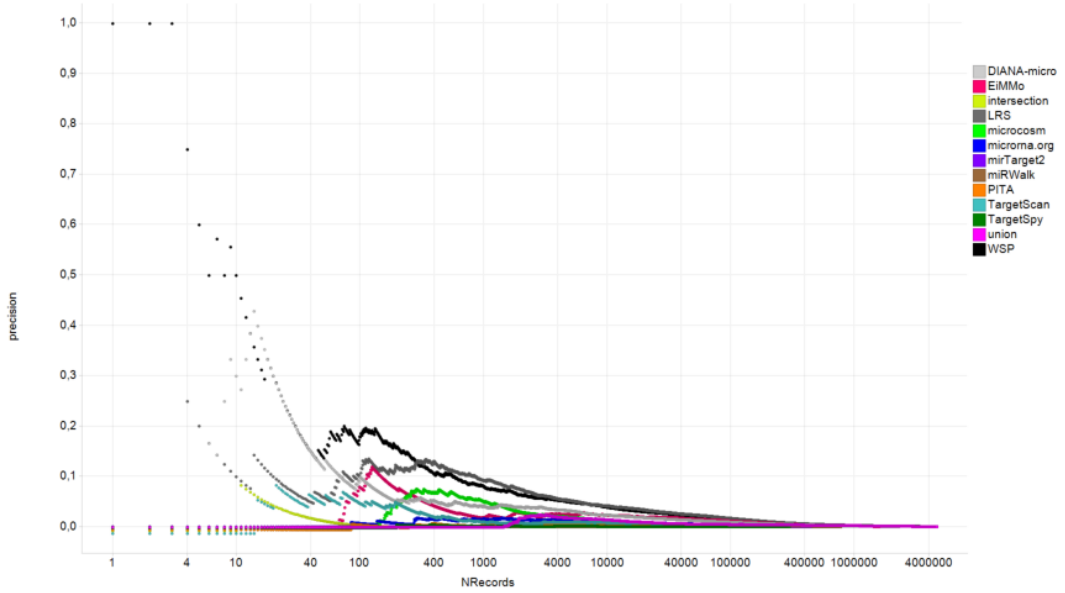


Figura 3.9: Curvas de precisión para métodos predictivos incluidos en este trabajo junto con los nuevos algoritmos propuestos.

sido validadas experimentalmente contra el total de predicciones. Esta aproximación también está afectada por la dependencia de los valores de falsos positivos, que no siempre puede estimarse, pero en este caso, al no utilizarse el valor de verdaderos negativos, se minimiza el error por la pérdida de información.

3.3.3. Métodos propuestos

3.3.3.1. *Weighted Scoring by Precision (WSP)*

En este trabajo se ha propuesto el siguiente método, que combina las puntuaciones de cada interacción en diferentes bases de datos predictivas calculando la suma ponderada de las puntuaciones normalizadas de cada predicción. Estos pesos se han incluido a causa de que los métodos de cálculo de las puntuaciones de las distintas bases de datos pueden no tener la misma fiabilidad.

El método propuesto se realiza de la siguiente forma, esquematizada también en la Figura 3.10. Primero las interacciones de cada base de datos predictiva se ordenan de mejor a peor

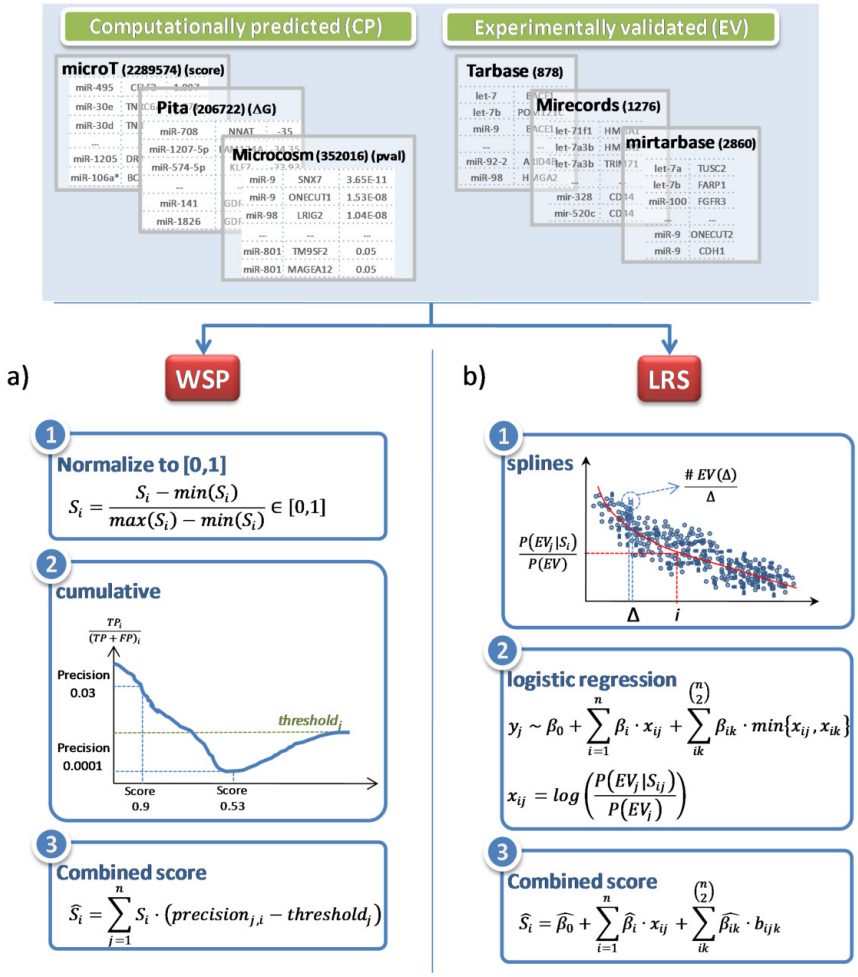


Figura 3.10: Esquema de funcionamiento de los dos métodos propuestos.

puntuación:

$$S_{ij} = \frac{S_{ij} - min(S_{ij})}{max(S_{ij}) - min(S_{ij})} \epsilon [0,1] \tag{3.8}$$

Entonces se calcula la precisión acumulada para cada una de las interacciones ordenadas hasta ese punto. De esta forma la precisión acumulada para determinada interacción sólo tiene en cuenta el número de verdaderos y falsos positivos desde la mejor interacción hasta ese punto. Se determinó la precisión umbral como la obtenida del cálculo de la última interacción, y para tener en cuenta la diferente fiabilidad en varias bases de datos predictivas, el valor de precisión es corregido restándole este valor. De esta forma, las interacciones con una precisión corregida menor que cero se podrían tomar como las que se comportan de forma similar a una predicción realizada al azar.

Por último, se calcula una puntuación para cada interacción, que integra todos estos cálculos previos para cada una de las bases de datos predictivas tomadas en consideración:

$$\hat{S}_j = \sum_{i=1}^{n} S_{ij} * (Precision_{ij} - Umbral_i) \tag{3.9}$$

Con este método, las interacciones con una puntuación combinada alta serán aquellas que, o bien han obtenido una puntuación muy alta en una base de datos, o han obtenido puntuaciones buenas en muchas de las bases de datos utilizadas. De esta forma, al tener en cuenta la precisión, las puntuaciones altas también significan mayor probabilidad de que se puedan validar experimentalmente. Si una interacción no ha sido correctamente puntuada en una de las bases de datos, muy probablemente esto no ocurra en las demás, y con la ponderación de las puntuaciones este efecto en la puntuación final se minimiza.

3.3.3.2. *Logistic Regression combined Scoring (LRS)*

En este estudio además del método *WSP*, se ha incluido también otro método de combinación de bases de datos de predicciones muy similar desarrollado por colaboradores de este trabajo. En esta aproximación se asume que cuanto más alta sea la probabilidad de que una interacción se pueda validar experimentalmente, más fiable es. En este método, descrito en la Figura 3.10, primero se determina la probabilidad de cada interacción de cada una de las bases de datos predictivas de haber sido experimentalmente validada, y entonces las combina para otorgar a cada interacción una única probabilidad, utilizando un modelo de regresión logística.

Primero, las interacciones de cada base de datos son ordenadas por sus puntuaciones, de mejor a peor. Entonces, las interacciones se agrupan por puntuación, y para cada grupo se calcula el ratio entre el número de interacciones experimentalmente validadas dentro del grupo con respecto al tamaño total del grupo. Estos ratios se interpolan creando una curva suavizada. Finalmente, se ajusta una regresión logística utilizando las puntuaciones extraídas de las curvas suavizadas, teniendo en cuenta que las mismas interacciones pueden tener diferentes puntuaciones en las distintas bases de datos predictivas. Las probabilidades devueltas por la regresión logística en las diferentes bases de datos se combinan entonces para crear las puntuaciones finales.

3.3.4. Desarrollo de aplicación web *m³RNA*

Las bases de datos combinadas obtenidas con estos dos métodos han sido incluidas en una aplicación web. La aplicación consta de una librería escrita en *Ruby*, que realiza consultas a una base de datos implementada en *postgreSQL*. Se pueden utilizar diversos tipos de identificadores tanto de micro ARNs como de genes, ya que la librería internamente los traduce a los formatos internos, *Ensembl* y *miRBase*. Esta librería está conectada con un servicio web *SOAP*, para proporcionar acceso programático para operaciones de lectura de datos. Los usuarios pueden acceder a la información por esta vía proporcionando un identificador de organismo y una lista de micro ARNs y/o genes. Los resultados se devuelven en forma de tabla, que contiene información de las bases de datos combinadas, datos de las bases de datos experimentales y predictivas utilizadas, en forma de puntuaciones normalizadas y datos sobre la precisión, y datos descriptivos acerca de los micro ARNs y genes utilizados en la consulta. Por encima de todo estos se ha construido una sencilla aplicación web con el framework *Ruby on Rails*.

3.3.5. Caso de uso

Una utilidad directa de esta herramienta es la de encontrar posibles micro ARNs que regulen la acción de determinado gen para entonces realizar una validación experimental. En la publicación de *Chen et al.*[45] se realiza una validación experimental de que el micro ARN *hsa-miR-30d-5p* inhibe la expresión del gen *CCNE2* mediante su unión en la región *3'UTR* del mismo realizando ensayos de genes reportero luciferasa y de inmunoprecipitación de ARN.

Si se quisieran encontrar posibles dianas para el gen *CCNE2* en el supuesto de no haber sido validado experimentalmente aún, a través de la web de la herramienta presentada, y seleccionando en el formulario el organismo adecuado (*Homo sapiens*) e introduciendo el nombre del gen en la caja de entrada, se obtendría una tabla de resultados de posibles interacciones ordenadas según la puntuación del algoritmo *WSP*. De entre los diez primeros resultados, cinco pertenecen a la familia de los micro ARN *miR-30*, siendo *hsa-miR-30d-5p* el cuarto, con una puntuación muy buena, como puede observarse en la Figura 3.11. El algoritmo *LRS* en este caso también corroboraría esta predicción, otorgando la mejor puntuación a esta posible interacción. Esta prueba se ha realizado no permitiendo la inclusión de la interacción entre el gen *CCNE2* y el micro ARN *hsa-miR-30d-5p* en las bases de datos de interacciones experimentalmente

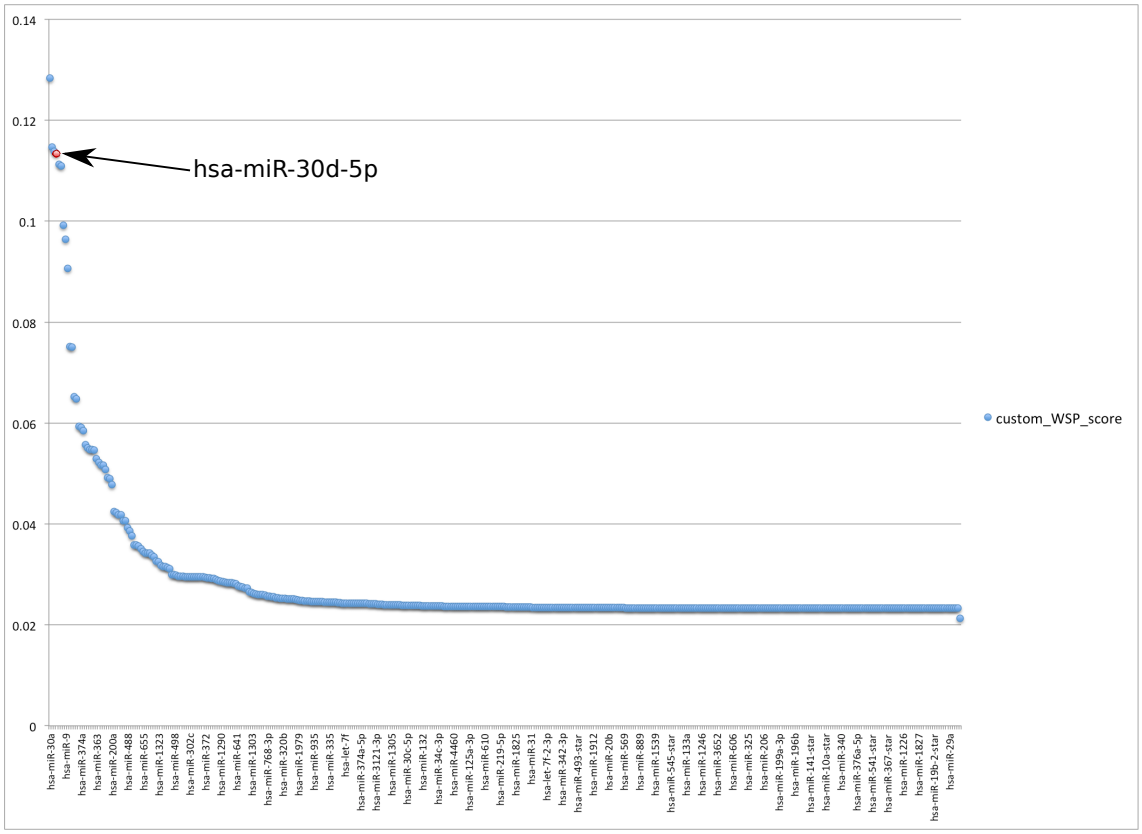


Figura 3.11: Gráfica de resultados del algoritmo *WSP* para el gen *CCNE2*.

validadas, para no sesgar el resultado, y asemejarlo a un caso real.

3.3.6. Resultados

Se ha utilizado un test hipergeométrico para poder medir la fiabilidad de las bases de datos predictivas. Los resultados de este test pueden verse en la Tabla 3.4, incluyendo también las bases de datos generadas por los métodos descritos en este trabajo. El valor z calculado para cada una de estas bases de datos es una medida de enriquecimiento en interacciones experimentalmente validadas. Ya que los métodos desarrollados en este trabajo combinan todas estas bases de datos de predicciones, es lógico que sean las mejor puntuadas en cuanto a este valor z, ya que contienen todas las interacciones predichas que han podido validarse del resto de bases de datos. *TargetScan* es de las bases de datos incluidas más pequeñas, pero contiene la mayor proporción de interacciones que han podido validarse experimentalmente. A pesar de esto, el valor z asignado a esta base de datos es bastante pobre. Por lo tanto, utilizar esta puntuación

como única medida de fiabilidad en las predicciones puede llevarnos a un error. Esta medida es fiable únicamente utilizando bases de datos de un tamaño similar.

Para poder comparar los resultados de los métodos propuestos, se han utilizado los métodos de evaluación propuestos anteriormente, las curvas *ROC* y las curvas de precisión. Además, hemos incluido otros dos métodos de combinación de bases de datos muy simples, la unión y la intersección. En la Figura [curva ROC] se pueden apreciar las curvas *ROC* para todas las bases de datos predictivos y los métodos propuestos. Se puede observar que la intersección es el método con un área bajo la curva mayor, pero esto es así debido a su minúsculo tamaño, contando sólo con 117 interacciones. Del resto de bases de datos, los dos métodos propuestos tienen un área bajo la curva igual o mejor que el resto de bases de datos de predicción tomadas en consideración, quedando *EiMMo* igualada con el método *WSP*. Es importante recalcar la diferencia de tamaño de las diferentes bases de datos, que al igual que con la base de datos intersección puede hacer que la comparación pueda verse distorsionada.

Para limitar los efectos de la información ausente en falsos positivos y verdaderos negativos, se ha propuesto en este trabajo la curva de precisión. Se ha calculado para todas las bases de datos predictivas y métodos propuestos, y aparece representada en la Figura [curva precision]. Como primera diferencia aparece que la base de datos de intersección, que aparecía con una mayor área bajo la curva utilizando curvas *ROC*, desaparece de entre las mejores. Es considerable también la diferencia entre las curvas de la base de datos *EiMMo* y el método *WSP*, que aparecían con la misma área bajo la curva en las *ROC*, y en cambio en este tipo de comparación queda muy por debajo. En la gráfica se puede apreciar que los dos métodos propuestos mejoran al resto, y se comportan de un modo muy similar exceptuando las primeras 400 interacciones aproximadamente. Esas diferencias constituyen un porcentaje muy bajo de las interacciones de los dos métodos, en torno al 0.01 %. En la experiencia a la hora de validar interacciones, no sólo es importante una buena puntuación de un determinado método, sino que también hay que tener en cuenta la posición de la interacción dentro del total de las interacciones ordenadas. De esta forma, los dos métodos propuestos son perfectamente compatibles, pudiendo elegir para validación las interacciones que aparezcan en lo más alto de la lista en ambos métodos.

Por último, todos los datos de las bases de datos tanto predictivas como experimentales, y de los dos métodos propuestos, están accesibles a través del portal desarrollado para tal efecto. Utilizando el buscador, aparecerá una lista de interacciones que puede ordenarse por diferentes valores, en las que aparece su presencia o ausencia en las bases de datos experimentales inclui-

das, los valores de predicción normalizados, así como la precisión y la precisión corregida de cada una de las bases de datos predictivas, así como para los dos métodos propuestos. Se pueden realizar búsquedas dentro de la tabla y consultar información de los genes y micro ARNs que aparecen en los resultados de la búsqueda, además de poder descargar todas las interacciones de una consulta en forma de fichero tabulado.

3.4. Una herramienta para buscar experimentos transcriptómicos similares en el contexto del reposicionamiento de fármacos (*NFFinder*)

En el proceso de desarrollo de nuevos fármacos se invierte una cantidad ingente de recursos y tiempo, llegando a pasar más de quince años hasta poder llegar a comercializarlos. Los compuestos en una fase inicial o preclínica son probados en cultivos celulares y animales para estudiar cómo actúa el organismo ante la molécula (farmacocinética), y cómo actúa la molécula frente al organismo (farmacodinámica). De esta manera se puede estudiar la posible toxicidad del compuesto o su forma de absorción y degradación por el organismo. Una vez superada esta fase, si el compuesto candidato es idóneo, pasa a fase clínica, con una primera fase inicial en la que se utiliza un reducido número de personas, donde se determina la posología y los posibles efectos secundarios, entre otros. Gran cantidad de compuestos candidato son rechazados en estas fases, sin llegar a fases clínicas más avanzadas o lanzarse al mercado, habiéndose invertido mucho tiempo y dinero en ellos. Las empresas farmacéuticas almacenan toda esta información sobre los compuestos, posibles dianas, efectos adversos, etc. Debido a los altos costes en este tipo de desarrollos, las farmacéuticas normalmente utilizan también otro tipo de estrategias complementarias. Una de ellas es el reposicionamiento de fármacos, que es básicamente la utilización de compuestos ya conocidos en enfermedades para las cuales no habían sido diseñados inicialmente[71]. De esta forma se pueden reducir los costes y el tiempo invertido hasta llegar a las fases de pruebas clínicas.

Gran parte de los fármacos reposicionados existentes lo fueron por accidente[21], por ejemplo al observar en fases clínicas efectos secundarios que podrían ser terapéuticos para otras enfermedades. Pero gracias a las nuevas tecnologías se ha pretendido evaluar todos estos compuestos descartados frente a nuevas dianas en busca de tratamientos. Existen principalmente dos estrategias en este sentido, las basadas en compuestos y las basadas en enfermedades[67]. Las basadas en compuestos intentan buscar similitudes a nivel químico, estructural o de actividad con compuestos ya utilizados para tratar determinada patología de interés. Las basadas en enfermedades buscan enfermedades con signos patológicos similares para intentar utilizar el mismo tipo de compuestos como tratamiento. Lo más efectivo suele ser combinar ambas estrategias.

Existe una forma alternativa de reposicionamiento a la hora de trabajar con enfermedades

de origen genético. Este tipo de enfermedades suele mostrar patrones anormales de expresión génica, que pueden investigarse para intentar revertirlos mediante fármacos. De esta forma se pueden aprovechar los datos presentes en bases de datos de expresión relacionados con compuestos y/o enfermedades para encontrar patrones de expresión similares e intentar hacer nuevas relaciones que desemboquen en un reposicionamiento. Gracias al auge en las tecnologías de secuenciación, se han creado repositorios públicos en los que almacenar experimentos de expresión de un modo muy homogéneo. En la sección 1.1.5.2.1 aparecen las principales bases de datos creadas en este sentido, siendo las más importantes *GEO* de *NCBI* y *ArrayExpress* de *Embl-EBI*.

En el caso de *Connectivity Map*[149] del *Broad Institute*, aparte de albergar datos de expresión, contiene una herramienta que permite realizar consultas contra sus datos, y de esta manera poder realizar comparaciones. Para llevar a cabo estas comparaciones utiliza un algoritmo de reconocimiento de patrones basado en el de *GSEA*[273]. De esta forma, y al utilizarse en los perfiles de expresión de esta base de datos cultivos de células humanas tratadas con diferentes compuestos, se pueden buscar conexiones funcionales entre estos compuestos, enfermedades y genes.

Tomando como base la idea de *Connectivity Map* de comparación de perfiles surgió la idea de extender este tipo de comparaciones a otras bases de datos de experimentos. De esta forma surge *NFFinder*, una herramienta web para poder crear hipótesis en reposicionamiento de fármacos basado en comparación de perfiles de expresión. Esta herramienta permite, a partir de la introducción de los genes diferencialmente expresados de un experimento, poder realizar consultas contra una base de datos que contiene datos de *GEO*, *CMap* y *DrugMatrix*, realizando comparaciones de firmas de expresión con un método basado en el utilizado en la herramienta *MARQ*[291]. Los experimentos extraídos de estas bases de datos han sido procesados utilizando herramientas de minería de texto como *MetaMap*[14], extrayendo información relacionada con fármacos, enfermedades y científicos para poder filtrar los resultados de una forma mucho más específica.

3.4.1. Construcción y comparación de perfiles

NFFinder genera una base de datos interna a partir de las miles de firmas de expresión génica de todos los experimentos extraidos de *GEO*, *CMap* y *DrugMatrix*. Estas firmas se generan

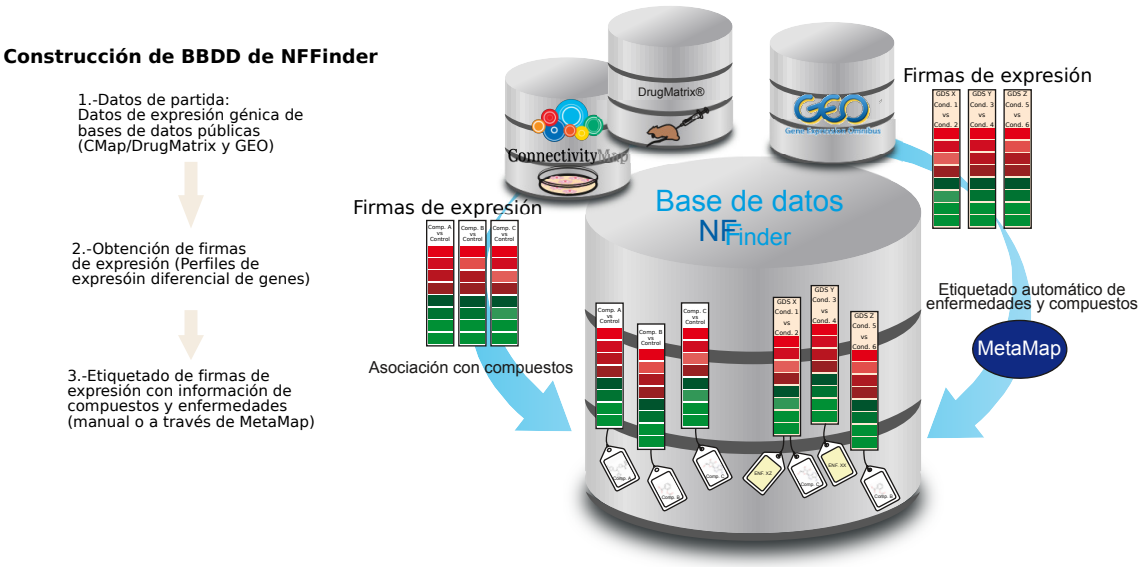


Figura 3.12: Esquema de construcción de la base de datos y etiquetado de firmas de *NFFinder*.

en base a la comparación de diferentes muestras de estas bases de datos utilizando *Limma*[240] para hacer un análisis de expresión diferencial. Una vez hecho esto, si existe un suficiente número de réplicas, los genes se ordenan por el valor *t* procedente de *Limma*, y si no, se ordenan por los valores de cambio de expresión entre las condiciones. En el caso de *CMap* y *DrugMatrix*, es trivial realizar estas comparaciones, ya que en estas bases de datos los compuestos utilizados fueron comparados con experimentos control. De esta manera, todas las comparaciones son razonables. No ocurre lo mismo con los datos de *GEO*. Los datos utilizados de esta plataforma son aquellos que han sido revisados manualmente e introducidos en los llamados *GEO DataSets* (*GDS*). Estos contienen información adicional sobre los factores presentes en el experimento, originalmente introducidos para poder realizar tareas de agrupamiento y comparación dentro de estos datos. Los factores pueden representar cualquier tipo de agrupamiento de muestras, desde tiempos, tipo de células, cultivo o tejido, compuesto utilizado, etc, y son los que se utilizan para comparar las muestras y poder crear perfiles de expresión. Pero no todas las comparaciones tienen sentido. En este trabajo se buscan primero palabras clave para encontrar muestras asignadas como control, y poder realizar comparaciones del resto de muestras contra estas. Si no se encuentra, se comparan todas las posibles condiciones, pudiendo aparecer comparaciones sin sentido. Estas comparaciones no pueden ser filtradas de forma automática, por lo que es tarea del usuario hacer esto a la hora de realizar una interpretación de los resultados. La dirección de la comparación en caso de no encontrar un control, o no identificar una serie de tiempos, se realiza también de forma arbitraria. Todos los resultados de las firmas de perfiles se almacenan en una base de datos local. El proceso de generación de perfiles y el posterior procesamiento de

las firmas para añadir distintos tipos de metadatos aparece esquematizado en la Figura 3.12.

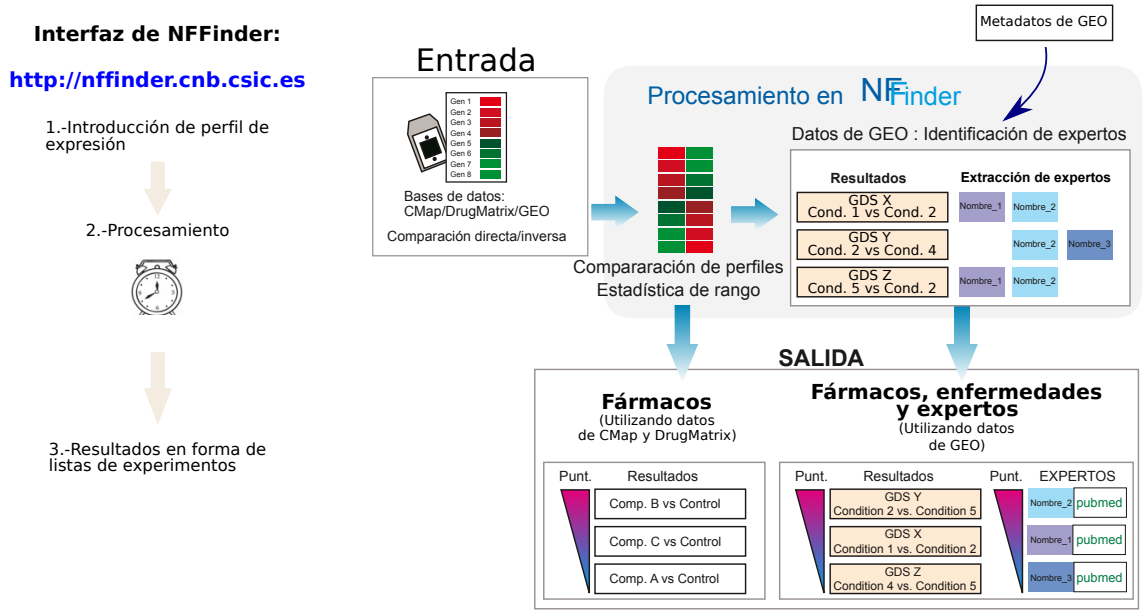


Figura 3.13: Esquema de funcionamiento de la comparación de perfiles en *NFFinder*.

Una vez generada la base de datos de perfiles, a partir de una lista de genes sobreexpresados e inhibidos de un experimento, podemos realizar comparaciones para poder encontrar experimentos con una expresión similar, como puede observarse en la Figura 3.13. Estas comparaciones se realizan utilizando una metodología similar a la empleada en la herramienta *MARQ*[291], calculando una puntuación de similitud basada en la utilizada en *Connectivity Map*. Los genes de entrada primero se dividen en dos listas separando los genes sobreexpresados de los inhibidos. Cada una de estas dos listas recibe una puntuación al compararse con un perfil de expresión, premiando a los genes de estas listas que se encuentran también diferencialmente expresados de una forma significativa en el perfil contra el que se compara. De esta forma los perfiles comparados son más parecidos si sus genes diferencialmente expresados son similares, recibiendo así mejores puntuaciones. Para calcular estos valores de puntuación sólo se pueden utilizar los genes de entrada que también aparezcan en la firma. Los genes reciben un peso en base a la posición en la firma, que depende de la distancia de ese gen al centro de la firma:

$$d_i \in [0,1] : d_i = \left| \frac{pos\_firma_i - \frac{Ng\_firma}{2}}{\frac{Ng\_firma}{2}} \right| \tag{3.10}$$

Gracias a esta distancia se pueden calcular unos pesos $peso_i = peso(d_i)$ utilizando la siguiente

función de peso:

$$peso(d_i) = \alpha\beta_0 d_i + (1-\alpha) * e^{\frac{\beta_1 d_i}{\beta_1}} \tag{3.11}$$

donde $\alpha = 0.5$, $\beta_0 = 0.3$ y $\beta_1 = 50$. Los genes que no aparecen en la firma se penalizan de la siguiente manera:

$$penalizacion = Ng\_no\_firma * peso(0.2) \tag{3.12}$$

También se calculan puntuaciones parciales para cada elemento de las listas de genes sobreexpresados e inhibidos por separado:

$$Punt\_parc_i^+ = \frac{\sum_{j=1}^{i} peso_j}{\sum_{j=1}^{Ng\_entrada} peso_j + penalizacion} - \frac{pos\_firma_i}{Ng\_firma} \tag{3.13}$$

$$Punt\_parc_i^- = \frac{pos\_firma_i}{Ng\_firma} - \frac{\sum_{j=1}^{i-1} peso_j + penalizacion}{\sum_{j=1}^{Ng\_entrada} peso_j + penalizacion} \tag{3.14}$$

Por último se calcula una puntuación final para cada una de las listas, que es el máximo de todas las calculadas:

$$Punt^+ = max(Punt\_parc_i^+) \tag{3.15}$$

$$Punt^- = -max(Punt\_parc_i^-) \tag{3.16}$$

Finalmente se calcula una puntuación global, que puede ser 0 si las dos puntuaciones parciales tienen el mismo signo y no son 0, o $Punt^+ - Punt^-$ si tienen diferentes signos. El signo de esta puntuación final indica si la relación entre la entrada y la firma es directa o inversa, es decir, si la expresión de la firma contra la que se compara la entrada es similar, o por el contrario es totalmente opuesta. En *NFFinder* se selecciona el tipo de relación buscada al inicio, por lo que se filtran todos los resultados de comparación del sentido contrario al indicado. Una vez calculadas todas las puntuaciones entre la entrada y las firmas de la base de datos, las mismas se escalan entre 0 y 100, siendo 100 la mejor puntuación, que correspondería a la firma más similar en el caso de una comparación directa, y la más opuesta en el caso de una comparación inversa.

La significación estadística de cada puntuación se calcula utilizando permutaciones aleatorias de listas de genes de entrada del mismo tamaño que la original y viendo la fracción de las que generan una puntuación mejor que la lista original. Posteriormente el valor p generado se corrige utilizando una corrección *FDR*[27].

3.4.2. Etiquetado de firmas

Las firmas de expresión de las bases de datos de *Connectivity Map* y *DrugMatrix* se asociaron de una forma directa a términos de fármacos y compuestos, ya que en las bases de datos de origen ya existen esas asociaciones. En el caso de *GEO* es más complicado debido a que esa información no está asociada directamente. Primero se extraen las descripciones de los *GDS* a los que pertenecen las firmas de expresión, y posteriormente se analizan utilizando una herramienta de análisis de lenguaje natural para términos biomédicos, *MetaMap*[15]. Este programa consigue extraer términos asociados a los existentes en *UMLS*[35]. Se realizó una búsqueda manual de términos *UMLS* asociados a compuestos químicos, fármacos y enfermedades, filtrando entonces los resultados de la ejecución del programa *MetaMap* con la descripción del *GDS* por estos términos. Una vez hecho esto, todas las firmas asociadas a un *GDS* son marcadas con los términos biomédicos resultantes de la ejecución de *MetaMap*, separando por un lado los términos asociados con compuestos químicos y fármacos, y por otro los términos asociados a enfermedades.

NFFinder además de generar una asociación de firmas con compuestos y enfermedades para poder así hacer relaciones entre fenotipos a través de la comparación entre diferentes firmas, también contiene asociaciones de las firmas de expresión con investigadores expertos. Los *GDS* de *GEO* contienen en su descripción un apartado con las citas de los artículos relacionados con esos datos, cuyos autores serían los expertos relacionados. De esta forma la relación entre las firmas de expresión contenidas en *GEO* y los expertos es directa. A la hora de comparar una lista de genes de entrada con la base de datos de perfiles de *NFFinder*, se genera una puntuación para cada experto, que es básicamente el número de experimentos asociados a firmas de expresión similares a la de entrada.

3.4.3. Caso de uso

NFFinder se ha desarrollado en el contexto de la investigación acerca de la neurofibromatosis, que es un grupo de enfermedades autosomales dominantes causadas por deficiencias en los genes de la neurofibromina. En el sistema nervioso periférico, la neurofibromatosis se manifiesta en forma de tumores benignos que pueden degenerar en tumores malignos de la vaina del nervio periférico (*MPNST*). La enfermedad más común de este tipo, NF1, ocure en 1:3000 na-

cimientos [260] y se considera una enfermedad rara. Como ejemplo para ilustrar la aplicación, se pretende buscar compuestos para revertir el fenotipo de líneas celulares *ST88-14* derivadas de tumores *MPNST*. Para poder encontrar una lista de genes diferencialmente expresados, se utilizaron resultados de análisis de *microarrays* comparando tumores *MPNST* contra células de *Schwann* normales[276].

En una primera aproximación se realizó una búsqueda inversa frente a *CMAP* y *DrugMatrix*. Como resultado se obtuvieron 775 perfiles con un total de 391 compuestos. Los 30 fármacos con mayor puntuación junto con los 10 más abundantes, formando un conjunto de 32 compuestos, están relacionados con tratamientos contra el cáncer (56 %), desórdenes neurológicos (12 %), problemas de la piel (12 %) y neoplasias benignas (3 %). El 40 % del total de fármacos relacionados con tratamientos contra el cáncer se utilizan para tratar diferentes tumores malignos del sistema nervioso, como por ejemplo la Tricostatina A, que es uno de los mejor puntuados, y que muestra efectividad en otros tipos de tumores como el cáncer de mama o el carcinoma de células escamosas gracias a su efecto de detener la proliferación celular y desencadenar la apoptosis[171].

Posteriormente se buscaron perfiles similares en *GEO* a partir del anterior perfil de entrada. En este caso de entre los resultados existe un porcentaje de entre un 20 % y un 30 % de perfiles entre los 200 con mejor puntuación cuyas comparaciones no tienen sentido. Del resto de comparaciones adecuadas, la mayor parte provienen del conjunto de datos *GDS2736* en el cual también se analiza la expresión de tumores *MPNST*. De esta forma se puede encontrar información adicional incluyendo publicaciones e investigadores trabajando en experimentos similares, que permitan de una forma más indirecta poder añadir más información acerca del perfil de entrada del programa.

3.4.4. Implementación

Las firmas de expresión procedentes de *GEO* fueron generadas utilizando el lenguaje de programación *R* que contiene la herramienta *GEOQuery*[58], utilizada para generar de forma automática las firmas de expresión a partir de los datos en crudo de los *GDS*, y posteriormente aplicar *Limma*[240] para realizar el análisis de expresión diferencial entre muestras.

Las listas de genes de entrada pueden introducirse en diversas nomenclaturas, habiéndose

utilizado las extensiones de *Python* de *BioMart*[263] para poder construir un diccionario interno y realizar las traducciones. De esta misma forma se han traducido las sondas de los genes de las firmas de expresión de *GEO* para poder realizar comparaciones entre diferentes plataformas. Adicionalmente se han incluido datos de interacciones validadas experimentalmente entre micro ARNs y ARN mensajero de las 4 bases de datos estudiadas en el trabajo de *m3RNA* (*miRWalk*, *miRecords*, *TarBase* y *miRTarBase*), y un diccionario para convertir entre diferentes nomenclaturas de micro ARNs procedente de *miRBase*, para poder realizar consultas a partir de micro ARNs que potencialmente regulan la expresión génica de una determinada firma.

NFFinder contiene una librería escrita en el lenguaje de programación *Python* que se encarga de generar, almacenar y acceder a la base de datos de firmas de expresión, para lo cual también se utiliza una base de datos *PostgreSQL*. Esta librería se ha escrito en forma de servicio web *REST* utilizando el *framework* web *Django* incluyendo además el *toolkit Django REST framework* para poder realizar consultas de un modo unificado, y poder hacer un manejo sencillo de las mismas, que se almacenan también en la base de datos, para de esta forma evitar calcular las mismas consultas varias veces. La inferfaz web utiliza *Javascript* para poder mostrar los trabajos y navegar a través de los mismos. Los trabajos de comparación de una lista de entrada con los perfiles de *NFFinder* se ejecutan en un *cluster* de computación que contiene 6 nodos con procesadores *Intel Xeon Quad-core*, para poder realizar las comparaciones en el mínimo tiempo posible.

3.4.5. Resultados

En este trabajo se presenta la herramienta *NFFinder*, la primera aplicación web de comparación de perfiles de expresión génica orientado al reposicionamiento de fármacos que utiliza datos extraídos de *GEO*, *Connectivity Map* y *DrugMatrix*. En total se han procesado 3254 *GEO GDS* para generar 16432 firmas de expresión génica. 6100 más fueron añadidas procedentes de *Connectivity Map*, y otras 5288 más de *DrugMatrix*, conformando un total de 27820 firmas diferentes. Éstas han sido asociadas con información de compuestos, fármacos y enfermedades con el propósito de poder establecer conexiones entre firmas de expresión similares y diferentes enfermedades, o conexiones entre firmas de expresión opuestas que permitan asociar determinado fármaco a la reversión de un fenotipo procedente de una patología. Además se ha implementado un método de asociación entre las listas de entrada con los estudios realizados por expertos, y de esta manera encontrar potenciales colaboradores a la hora de estudiar una determinada en-

fermedad o compuesto. Todo esto se ha presentado como una interfaz web *REST* que permite realizar consultas de forma programática, y que permite navegar por los resultados expuestos en formato *JSON* de una manera sencilla.

Existen otras herramientas para realizar este tipo de comparaciones, como *Connectivity Map*, que sirvió de inspiración para esta herramienta, o *Combinatorial Drug Assembler* (CDA)[153]. La ventaja sobre *Connectivity Map* es clara, se utiliza una metodología de comparación similar, pero la base de datos de firmas de expresión es mucho más grande en el caso de *NFFinder*, por lo que la posibilidad de encontrar perfiles más similares u opuestos es mucho mayor. La herramienta *CDA* parte de dos listas de genes, sobreexpresados e inhibidos, las cuales son procesadas para posteriormente realizar un análisis de enriquecimiento de rutas de señalización. *CDA* contiene una base de datos que parte de resultados de enriquecimiento de rutas de señalización y de fármacos procedentes de utilizar las firmas de expresión de *Connectivity Map*, y busca similitudes entre estos análisis de enriquecimiento y el llevado a cabo a las listas de entrada. Finalmente se generan listas de fármacos asociados a perfiles con patrones de expresión similares. Esta herramienta realiza un análisis mucho más complejo, teniendo en cuenta las similitudes entre rutas metabólicas. Pero esta forma de realizar las comparaciones depende del conocimiento acerca de estas rutas, además de las asociaciones entre perfiles y fármacos o enfermedades, por lo que el grado de error puede ser mayor, o incluso puede que se estén obviando mecanismos de señalización no conocidos hasta el momento. En este caso además se utilizan solamente datos de expresión de *Connectivity Map*. Estos sistemas asimismo requieren de un registro previo al análisis, no necesario en *NFFinder*, demorando de esta forma el tiempo necesario en realizar una consulta. La información proporcionada como resultados en *CDA* es además incompleta y confusa, mientras que los resultados de *NFFinder* contienen la lista de firmas similares ordenadas por grado de similitud.

3.5. Una visión unificada, enriquecida e interactiva de la información sobre macromoléculas (*3DBionotes*)

Los avances en microscopía en los últimos años están permitiendo conocer mucho más a fondo las estructuras de gran cantidad de biomoléculas. En concreto, el campo de la microscopía electrónica ha vivido una revolución gracias a los avances en la criomicroscopía electrónica, que con el uso de nuevos detectores que no necesitan convertir los electrones en fotones para realizar la detección[20], junto con la aparición de una nueva generación de *software* para mejorar la reconstrucción de los volúmenes en la que se incluye *EMAN*[278], *RELION*[250] y *Xmipp*[59] entre otros, ha permitido reportar gran cantidad de estructuras a una resolución de pocos angstroms[146]. De esta forma, cada día aumenta el número de volúmenes reportados en bases de datos como *EMDataBank*[151] y estructuras tridimensionales en bases de datos como *wwPDB*[29].

En paralelo a esto, la cantidad de información a nivel de secuencias de genes y proteínas se ha disparado en los últimos años debido a las diferentes técnicas de alto rendimiento que se han desarrollado. Gran parte de esta información está centralizada en sitios como *Ensembl*[315] y *RefSeq*[215] para genes o *Uniprot*[305] para proteínas. Otro tipo de información más específica aparece dispersa en un gran número de bases de datos públicas, con datos acerca de, por ejemplo, mutaciones relacionadas con enfermedades, regiones relacionadas con alguna característica estructural concreta, etc.

La integración de estos distintos tipos de información es fundamental para ahondar en el conocimiento de las proteínas. Pero hasta hace poco tiempo, hacer esto de una forma sistemática era algo menos que imposible. La información estructural ha aparecido tradicionalmente almacenada en el formato de fichero *PDB*[30], pero aunque se sigue utilizando hoy en día, quedó obsoleto debido a diversas limitaciones, y apareció el formato *PDBx/mmCIF* que es el formato estándar hoy en día. En estos ficheros la información estructural aparece asociada a cadenas de aminoácidos que no tienen por qué corresponder con las secuencias consenso de las bases de datos de secuencias, o pueden contener regiones mal mapeadas o directamente sin información. De esta forma, para poder mapear información de secuencias en estructuras con estos formatos era necesario realizar un alineamiento entre la secuencia y la cadena del fichero de estructura. Hacer esto de una manera sistemática requiere de una capacidad de computación inmensa, ya que en la actualidad existen en la base de datos *wwPDB*[29] 131564 estructuras

depositadas, y la base de datos de proteínas revisadas de *Uniprot*, *Swiss-Prot*, contiene 553655 entradas actualmente. Afortunadamente, han surgido recientemente recursos como *SIFTS*[293] que han realizado este trabajo de correspondencia de un modo semiautomático, haciendo uso de las referencias cruzadas con *Uniprot* generadas en el depósito de estructuras en *wwPDB* para identificar el organismo del que procede la estructura y tener una correspondencia entre identificadores de estructura y de proteínas, y realizando posteriormente alineamientos de los segmentos conectados en las secuencias de las estructuras (de los extremos N-terminal a C-terminal) con las secuencias de las proteínas. De esta forma se ha abierto una oportunidad para poder hacer esta integración de una forma sencilla y sistemática, tarea abordada en el presente trabajo con la creación de *3DBionotes*, una herramienta web que permite visualizar estructuras, y seleccionar y visualizar diferentes anotaciones a nivel estructural o de secuencia de una forma sencilla.

3.5.1. Información estructural

Para poder visualizar toda la información estructural es necesario primeramente descargar las estructuras de *wwPDB* y los mapas de *EMDB*. Las estructuras procedentes de *wwPDB* se descargan en tiempo real con cada consulta realizada, y se mantienen en una carpeta en forma de caché para poder realizar consultas de forma más rápida. En el caso de los mapas de *EMDB*, es más complicado, ya que son ficheros que ocupan bastante espacio, del orden de cientos de megabytes, además de que tienen que ser remuestreados para poder visualizarlos de forma fluida por el navegador. Por esta razón, en un primer momento se descargó la lista completa de mapas de *EMDB*, y periódicamente se va actualizando con la descarga de las entradas nuevas. Después de su descarga, estos ficheros se transforman utilizando el paquete *Xmipp*[59].

3.5.2. Mapeo de estructura a secuencia y anotación

Antes de realizar mapeos entre estructuras y secuencias, el primer paso es obtener la relación de identificadores equivalentes en las diferentes bases de datos utilizadas. En este trabajo se relacionan tres tipos de datos, volúmenes de microscopía electrónica procedentes del *EMDataBank* (*EMDB*)[151], estructuras procedentes de *wwPDB* y secuencias de proteína de Uniprot. Para ello se han utilizado los servicios web *REST* de *EBI*[191] que permiten, a partir de urls,

obtener información acerca de estos mapeos en formato *JSON*. En concreto se utilizan tres servicios diferentes de estos servicios web, que son el servicio *fitted* que es básicamente el mapeo entre mapas de *EMDataBank* y estructuras de *wwPDB*, el servicio *uniprot* que permite relacionar los identificadores de *PDB* con las accesiones de *Uniprot*, y el servicio *best_structure*, que utiliza datos de *SIFTS* para dar una lista de las estructuras de *PDB* asociadas a proteínas de *Uniprot*. De esta forma, partiendo de cualquiera de estos tipos de identificadores, podemos llegar a obtener una correspondencia con los otros. Todos estos datos de correspondencias son almacenados en una base de datos local una vez realizada una primera consulta, para funcionar como una caché y minimizar el tiempo en el que se realizan.

Una vez obtenidas las correspondencias de identificadores, el primer paso para poder realizar el mapeo de estructura a secuencia consiste en utilizar el identificador de *PDB* de la estructura para acceder al servicio *SIFTS*. Este servicio proporciona un fichero de tipo *XML* por cada estructura, con información detallada de mapeo a nivel de residuo con varias bases de datos, entre ellas *Uniprot*. Estos ficheros *XML* pueden llegar a ocupar bastantes megabytes de espacio, y requerir de varios segundos para su procesamiento, por lo que hacer una gestión eficiente de su manejo es imprescindible. En esta herramienta se ha creado una base de datos donde almacenar los datos de mapeo a nivel de residuo entre los identificadores de *PDB* y *Uniprot*. La primera vez que se intenta acceder a los datos de mapeo de la base de datos de un identificador *PDB* determinado, estos no existen, por lo que el sistema descarga el fichero desde el servicio de *SIFTS*. Una vez descargado, genera una clave *MD5* del mismo, y se analiza para extraer la información de mapeo, que se almacena en la base de datos en formato *JSON*. En posteriores accesos el sistema puede utilizar la información almacenada localmente y ahorrar el tiempo de la descarga y procesamiento. Las entradas descargadas y almacenadas en la base de datos caducan después de un cierto tiempo, ya que la información proveniente de *SIFTS* puede modificarse. Si una entrada ha caducado, se descarga de nuevo el fichero *XML* de *SIFTS* y se compara su clave *MD5* con la almacenada previamente. Si coincide, los datos locales permanecen y el tiempo de caducidad se amplia. Si no coinciden, se procesa el fichero *XML* descargado y se reemplazan los datos de la entrada.

Por último, la información de anotaciones de secuencias proveniente de distintas fuentes de datos, tales como *dSysMap*[202], *BioMuta*[310], *Immune Epitope DB*[297], *PhosphoSitePlus*[117] o la propia *Uniprot*, requieren de un procesamiento distinto para cada una de ellas. En el caso de *Uniprot* y *dSysMap*, las anotaciones se extraen de los servicios web que proporcionan ambas bases de datos para tal efecto, permitiendo un acceso por identificador individual.

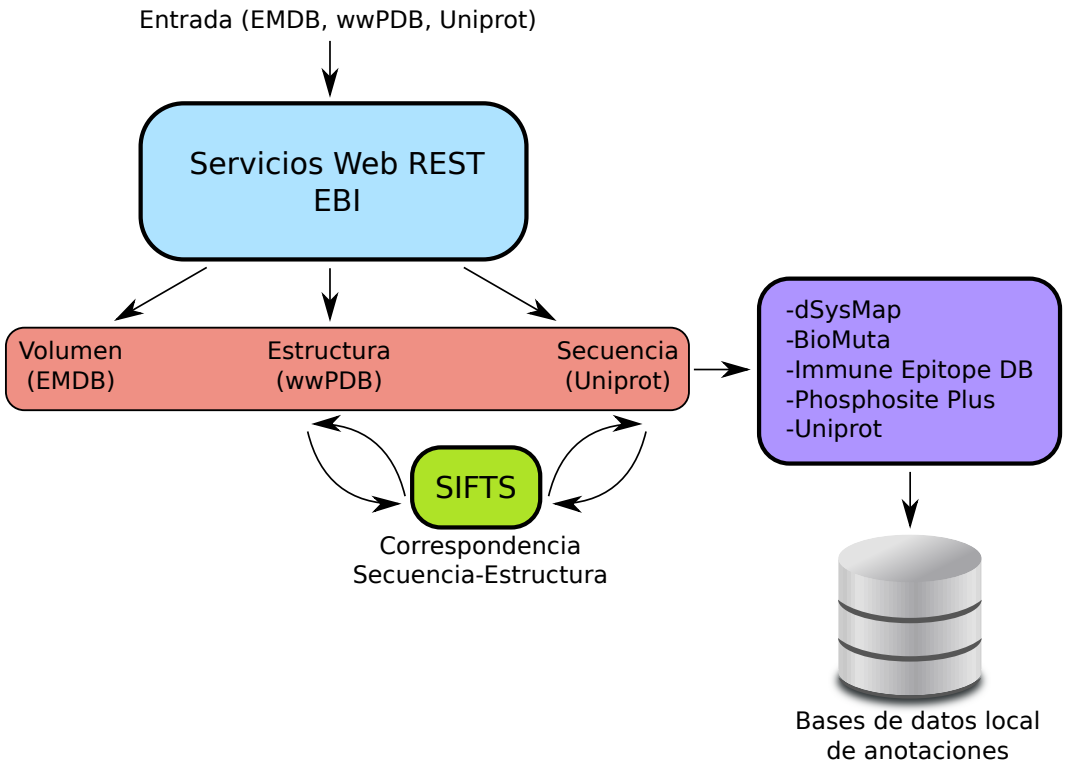


Figura 3.14: Esquema del origen y relación entre distintos tipos de identificadores e información de bases de datos estructuales y de secuencia presentes en *3DBionotes*.

Para *PhosphoSitePlus* y *BioMuta* es necesario descargar ficheros en formato tabulado que tienen que ser procesados. En el caso de *Immune Epitope DB* es necesario descargar una base de datos *MySQL*, extraer la información de la misma y procesarla. Toda esta información es posteriormente introducida en una base de datos local, en la que para cada identificador de proteína de *Uniprot* aparece una lista de anotaciones con información para cada una de ellas acerca de su base de datos de origen, tipo de anotación, coordenadas de comienzo y final, y un campo de descripción en el que se almacena en formato *JSON* un conjunto de datos adicionales para cada tipo de anotación, que es diferente para cada una de las bases de datos incluidas. De esta forma este diseño en la base de datos es extendible a la adición de nuevas fuentes de datos. Simplemente hay que convenir un formato de descripción en *JSON* para poder describir las nuevas anotaciones. Las anotaciones provenientes de *Uniprot* y *dSysMap*, al poder acceder a entradas individuales, permiten que se haga un acceso al vuelo, manteniendo posteriormente una caché de consultas para un acceso aún más rápido similar a la utilizada en el caso de los mapeos de estructuras de *SIFTS*. Todo este proceso de mapeo y obtención de información de bases de datos aparece esquematizado en la Figura 3.14.

3.5.3. Interfaz gráfica

En la herramienta *3DBionotes* es necesario visualizar la información estructural de los mapas de *EMDB* y estructuras de *wwPDB*, combinado con las anotaciones de secuencia de los alineamientos con *Uniprot*. Para llevar a cabo esta tarea se ha desarrollado un panel web dividido en cuatro partes.

Un pequeño panel superior muestra el menú de la aplicación, con acceso a una sección de ayuda, junto con el nombre de la consulta realizada, y un selector para poder elegir el mapeo entre estructura y secuencia más conveniente.

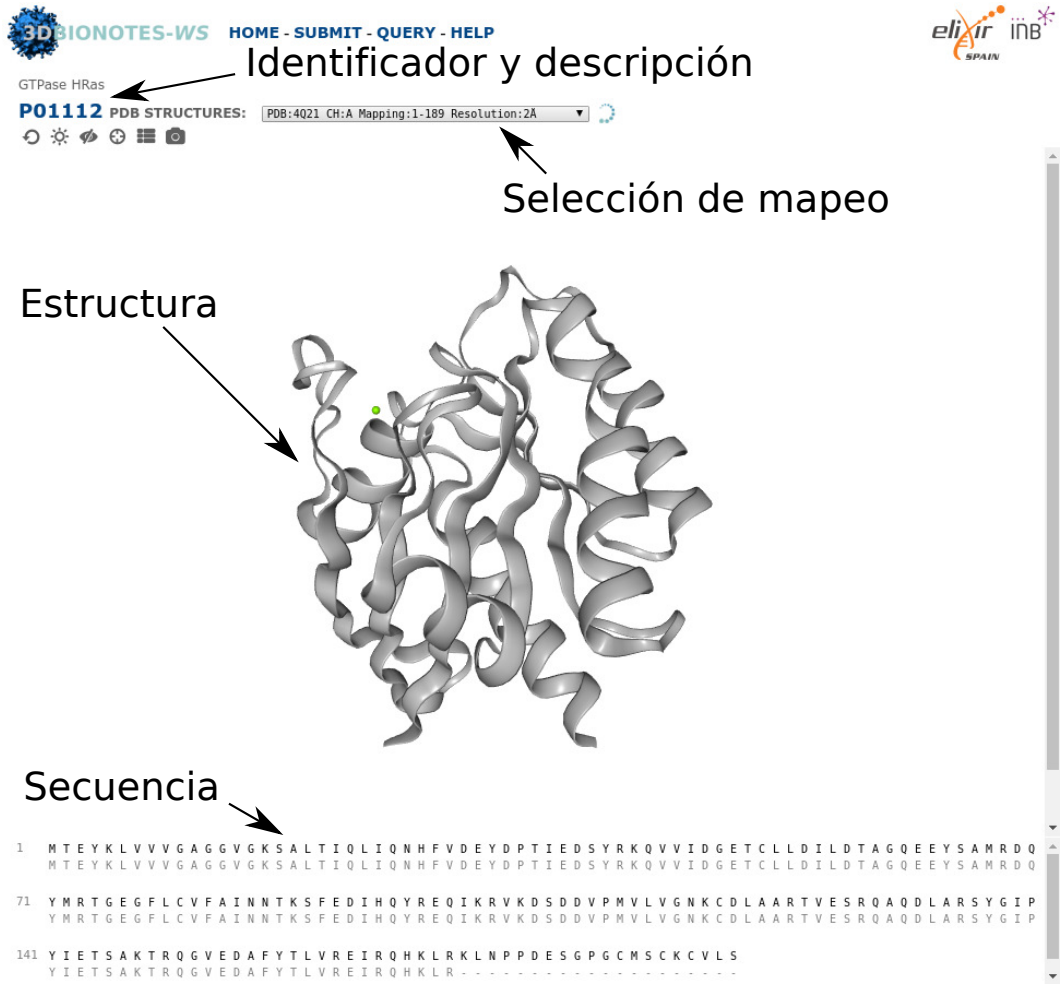


Figura 3.15: Paneles de visualización de estructura y alineamiento entre estructura y secuencia en un ejemplo utilizando *3DBionotes*.

El panel de la izquierda sería el que contiene la visualización de las estructuras. Para ello se ha utilizado el visualizador *JSmol*[103]. Este panel utiliza directamente los datos descargados de *wwPDB* y los mapas remuestreados de *EMDB* para la visualización. La visualización puede ser modificada, cambiando la cámara utilizando el ratón, o a través de los botones superiores, que permiten modificar el zoom de la cámara, visualizar la región más cercana a las anotaciones seleccionadas, reiniciar la visualización, o incluso tomar una instantánea para poder descargarla. Para comunicar información y cargar los datos de estructura en el visualizador ha sido necesario construir una pequeña librería para encapsular los comandos necesarios para tal efecto.

Un panel en la parte inferior de la pantalla contiene el alineamiento entre la cadena seleccionada de la estructura de *PDB* y la secuencia de *Uniprot*. La visualización de este alineamiento se ha llevado a cabo mediante el uso del *framework* de *JavaScript BioJS*[95], en concreto el paquete *Sequence*[94]. Se pueden seleccionar regiones del alineamiento, quedando estas coloreadas y marcadas también en el resto de paneles. Si en otro panel se selecciona una región, también aparecerá marcada aquí. Tanto el panel de visualización de estructuras como el de alineamiento entre la cadena seleccionada y la secuencia aparecen visualizados en la Figura 3.15.

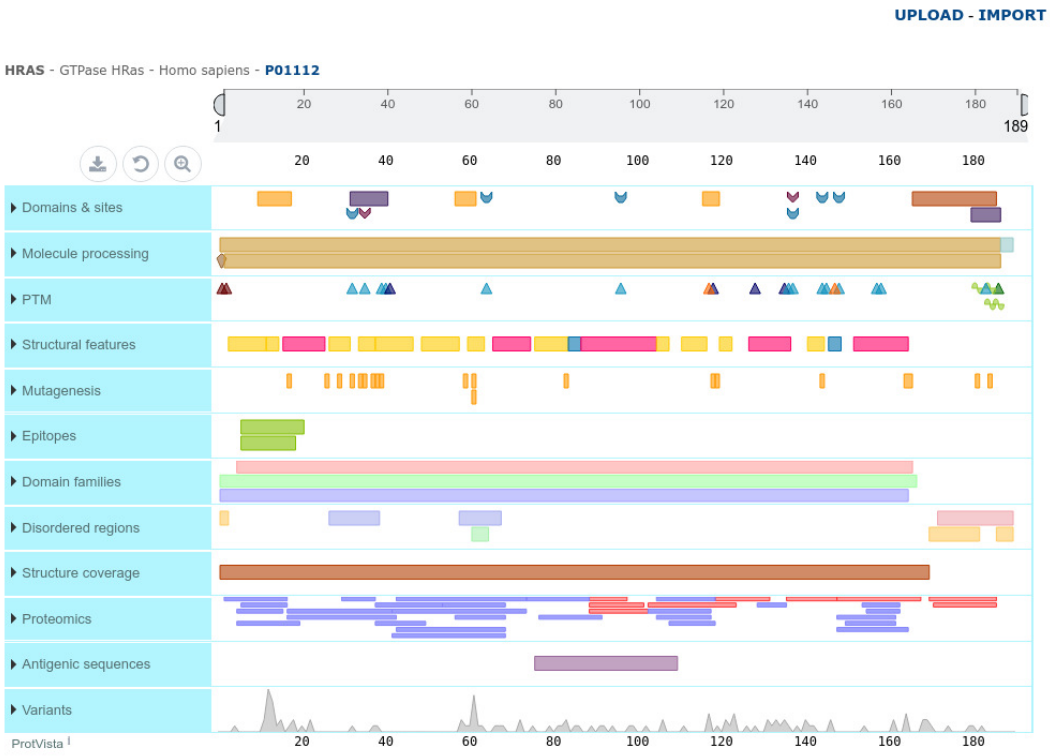


Figura 3.16: Ejemplo de panel de anotaciones en la aplicación *3DBionotes*.

El último panel, representado en la Figura 3.16, es el que contiene las anotaciones. Se ha utilizado en este caso una versión modificada de *ProtVista*[305], un visor de anotaciones que ha sido incluido como un componente de *BioJS*. Este visor utiliza pistas para visualizar diferentes tipos de información, pudiendo ver la coocurrencia espacial de las mismas de un modo sencillo. Cada pista puede ser expandida para mostrar información adicional de las anotaciones. También se puede realizar zoom para ver con más detalle las anotaciones de una determinada región. Estas anotaciones pueden ser seleccionadas para visualizar la información a nivel de alineamiento de secuencias o de forma visual con la estructura en los diferentes paneles.

Todos los paneles están relacionados entre sí de forma que toda la información quede sincronizada entre ellos, y quede realzada la información equivalente para todos los paneles.

3.5.4. Caso de uso

La proteína HRas GTPasa regula la división celular actuando como un interruptor molecular de la propagación de señales de los receptores de la membrana celular. En su forma activa está unida a un GTP, pudiendo a su ver activar numerosas rutas de transducción de señales. Cuando la molécula se desactiva se libera un fosfato y se mantiene unida al GDP. La familia de proteínas a las que pertenece es una familia de protooncogenes debido a su importancia en el proceso de la división celular.De hecho, mutaciones en este tipo de proteínas son la causa de un gran número de cánceres[136].

3DBionotes puede utilizarse para explorar la estructura y las anotaciones de esta proteína, codificada en *Uniprot* como *P01112*. En el ejemplo, a partir de este identificador se ha mapeado con la estructura de *PDB 4Q21*. Esta proteína tiene tres regiones en las que interactúa con la molécula *GTP/GDP*. En el panel de anotaciones, como puede observarse en la Figura 3.17, en la sección de variantes se puede observar una gran cantidad de mutaciones sobre todo en una de estas regiones, entre los aminoácidos 10 y 17, la mayoría de ellas asociadas a enfermedades. De esta forma se puede visualizar de una forma muy directa la importancia de estos sitios de interacción con *GTP/GDP* y su relación con enfermedades.

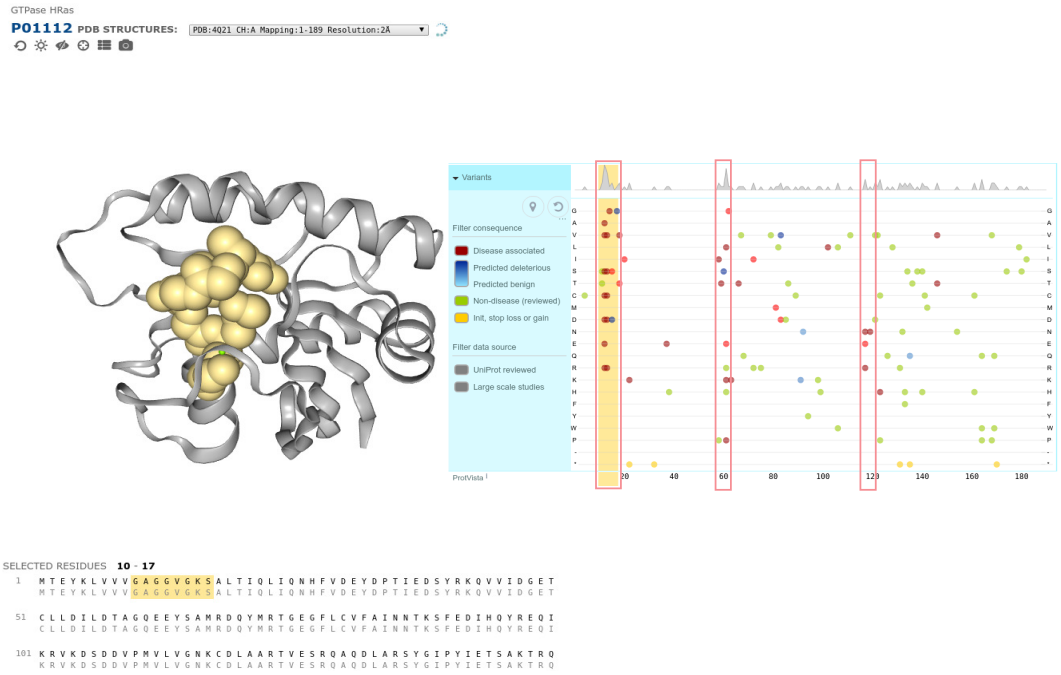


Figura 3.17: Visualización de la proteína HRas GTPasa, con las regiones de interacción con *GTP/GDP* marcadas.

3.5.5. Implementación

Se ha creado una librería escrita en *Ruby* para poder acceder y procesar a los datos de estructura, secuencia, mapeo y anotaciones. Todo esto ha sido contenido dentro del *framework* web *Ruby on Rails* para poder implementar un servicio web *REST* que centralice todas las consultas, que genera datos en formato *JSON* para poder realizar un procesamiento sencillo. Todos los datos de mapeo, secuencia y anotaciones han sido almacenados de la forma anteriormente expuesta en una base de datos *MySQL* utilizando campos de texto extensos con información *JSON* para tipos de datos complejos. Como se ha comentado anteriormente, no todos los datos están contenidos en la base de datos desde un inicio, sino que se van descargando y procesando bajo demanda, utilizando en gran medida la base de datos local como caché para agilizar el acceso a la información. Los datos de las estructuras *PDB* son almacenadas en una carpeta local según se van descargando, al igual que la información de *EMDB*.

La visualización web se ha creado también dentro del *framework Ruby on Rails*. Cada uno de los paneles de resultados es un *frame* distinto, que requiere de la utilización de mensajes asíncronos para su intercomunicación a través de la página web que los contiene, permitiendo una gran interactividad entre las diferentes visualizaciones. Se ha hecho un uso intensivo de

librerías *JavaScript* para las visualizaciones, y de esta forma utilizar en la manera de lo posible los recursos del ordenador que está generando la visualización y descargar al servidor de estas tareas.

3.5.6. Resultados

3DBionotes es una plataforma para integrar datos de diferentes fuentes biológicas presentada en el contexto de la biología estructural. De esta forma, se permite a los investigadores poder trabajar, con un mínimo esfuerzo y sin instalar ningún tipo de *software* adicional, con datos estructurales, manteniendo una visualización fluida tridimensional, solapando los diferentes tipos de información a nivel de secuencia encima de ella, para poder llegar a un mejor entendimiento de la función de estas estructuras y su rol en el metabolismo celular. Hasta el momento de la publicación no se ha encontrado ninguna herramienta con una funcionalidad similar. De hecho, esta herramienta ha sido enlazada desde la propia base de datos de *EMDB*, apareciendo un enlace dentro de cada entrada de mapa de microscopía de su sitio web.

El desarrollo de la herramienta ha sido marcado por su diseño fácilmente extensible. En el momento actual se está terminando de desarrollar una segunda versión de la herramienta, con nuevas bases de datos de información biológica, entre otras novedades. La creación de un marco genérico para las anotaciones, reflejado en un único tipo de entradas en la base de datos, es la principal razón de la sencillez de esta característica. Para añadir una fuente de datos simplemente habría que añadir un nuevo procesador de información para esa base de datos, y dependiendo de la posibilidad de descargar anotaciones de forma individual, o de descarga de toda la base de datos de golpe, habría que, o bien añadir esa fuente de datos como un tipo de anotación cacheable, o descargable en bloque. Habría que añadir también código fuente para incluir ese nuevo tipo de datos en el panel de anotaciones.

El diseño modular de los paneles de visualización también es otra de las características importantes de esta herramienta. Cada panel de visualización es totalmente independiente del resto, estando conectados todos por un sistema de paso de mensajes asíncrono. De esta forma, se podrían añadir nuevos paneles a la herramienta de una forma sencilla. O desarrollar un visor equivalente en otra tecnología y sustituirlo sin realizar cambios sustanciales en el código. De hecho, en una primera fase de desarrollo se implementó el panel de anotaciones con otra tecnología, y una vez aparecido *ProtVista* se realizó el cambio del panel.

3.6. Diseño de infraestructura y computación de alto rendimiento

El despliegue de las herramientas anteriormente explicadas de una forma eficiente requiere de conocimiento de diferentes sistemas y tecnologías. Poder gestionar los recursos de estos sistemas también es una gran ventaja a la hora de maximizar el rendimiento de estas aplicaciones. Existen muchas aproximaciones diferentes a la hora de gestionar los trabajos necesarios por estas herramientas web, pero la limitación de recursos obliga a descartar muchas de las soluciones más modernas y eficientes. En la actualidad, gran parte de las aplicaciones web desarrolladas por grandes empresas tienen un servidor *front-end* que recibe las peticiones, y permite redireccionarlas a una máquina con recursos libres que contenga un servidor web y gestione la petición. Gracias a la gestión de las peticiones, y a la existencia de un gran número de máquinas que puedan actuar de servidor web, permiten una respuesta inmediata incluso con un gran número de peticiones simultáneas. Este sistema de escalado horizontal permite, a costa de utilizar más recursos, evitar la saturación del servicio. De todas formas, no es posible comparar herramientas web de grandes empresas, que reciben millones de peticiones diarias, con herramientas realizadas para la investigación en un ámbito determinado, con un número mucho más limitado de peticiones. De hecho, la herramienta de este trabajo que más visitas recibe es *GeneCodis*, que registra entre 40 y 80 visitas diarias. Las técnicas de computación de alto rendimiento desarrolladas en los trabajos realizados son las siguientes.

3.6.1. Integración de computación en *cluster* en herramientas web

Las aplicaciones web, y el *software* en general, se puede dividir en dos partes, la parte de visualización y navegación, o *front-end*, y la parte de acceso a base de datos y procesamiento de información, o *back-end*. Es importante realizar una separación clara entre estas dos capas, ya que de hecho en la tecnología web parte de la carga de la visualización se genera del lado del cliente, el cual no debe sobrecargarse. En aplicaciones web de consulta a base de datos que no reciban una gran cantidad de accesos, una opción podría ser utilizar un único servidor para todo, incluyendo el servidor web y el servidor de bases de datos. Pero en aplicaciones más intensivas computacionalmente hablando, es imposible mantener toda la carga de procesamiento en un mismo servidor, ya que este puede llegar a saturarse y generar una caída de servicio.

Una opción podría ser realizar un escalado horizontal de la aplicación, replicando la misma en diferentes servidores que repartan la carga. Pero esta opción es muy costosa. Una alternativa, que es la que se ha desarrollado en varios de los trabajos de esta tesis, sería la de utilizar un servidor como *front-end*, con también parte del *back-end* situado en el mismo, como la encargada de gestionar el acceso a bases de datos y la gestión de trabajos. Este servidor está conectado con un *cluster* de computación, siendo capaz de enviar trabajos al mismo. La parte del *back-end* encargada de gestionar los trabajos, genera la información requerida como entrada para el proceso que se lanza después a la cola del *cluster*, realiza consultas al *software* encargado de la gestión de colas, que en este caso es *Open Grid Scheduler*, permite hacer seguimiento de los procesos enviados, y de esta manera actualizar en el gestor de trabajos de la herramienta la información de los mismos, hasta que estos finalizan. De esta forma se puede garantizar la ejecución de un trabajo dentro de la aplicación en el mínimo tiempo posible, evitando además el bloqueo de la herramienta. Las herramientas *GeneCodis* y *NFFinder* hacen uso de esta arquitectura, así como otras herramientas en las que se ha colaborado, como *3DEM Loupe*[213] o *Breaking-Cas*[216].

3.6.2. Plataformas virtualizadas

La utilización de un servidor para una única aplicación web puede tener como consecuencia que el servidor esté ocioso la mayor parte del tiempo. Con los avances de los últimos años en el ámbito de la virtualización, es posible particionar un servidor para poder aprovechar estos recursos ociosos. Existen diferentes tecnologías para hacer esto, como utilizar virtualización clásica a través de hipervisores, con sistemas operativos específicos que permitan una virtualización de tipo 1 (*Xen*[22], *VMWare ESXi*[204] o *Microsoft Hyper-V*[131]), o hipervisores de tipo 2 que se ubican por encima de un sistema operativo general (*Oracle VirtualBox*[306], *KVM*[139] o *VMWare Workstation*[274]), o utilizar tecnologías más modernas como los contenedores de aplicaciones (*LXC*[110] o *Docker*[193]). De estas dos soluciones, la única que permite un aislamiento total de los recursos de un servidor es la de máquinas virtuales. De esta forma en una sóla máquina potente podemos tener decenas de máquinas virtuales, cada una con los recursos básicos para poder mantener un servidor web que sólo se encargue del *front-end* y del lanzamiento de trabajos. En el caso de utilizar un hipervisor de tipo 2, se podrían mezclar en una misma máquina la presencia de máquinas virtuales y servicios comunes del sistema operativo nativo, que podrían estar instalados a partir de contenedores, como por ejemplo servidores de bases de datos. Si se utilizan hipervisores de tipo 1, como ha sido en este caso con la utilización del hipervisor de *VMWare ESXi*, no se podría hacer uso de recursos del sistema operativo nativo,

pero en cambio la eficiencia de las máquinas virtuales es mayor.

3.6.3. Almacenamiento compartido

A la hora de trabajar con *clusters* de computación, es fundamental mantener un espacio de almacenamiento común al que puedan acceder todos los nodos de procesamiento, y de esta manera evitar el coste adicional de transferir los ficheros entre discos de diferentes máquinas. Existen diferentes soluciones a nivel de *hardware* para hacer esto, como son los *Storage Area Networks (SAN)* o los *Network-Attached Storage (NAS)*. En el caso de los *SAN* se proporciona un almacenamiento en red a nivel de bloque, por lo que el sistema de ficheros queda del lado del cliente, mientras que los *NAS* proporcionan almacenamiento a nivel de ficheros, por lo que el propio servidor *NAS* se encarga del sistema de ficheros. Los sistemas *SAN* son mucho más costosos que los *NAS*, y también más eficientes, pero a la hora de compartir ficheros con varias máquinas a la vez, los *SAN* requieren de sistemas de ficheros especiales que producen un rendimiento final no tan alto. Por lo tanto en el caso de este trabajo, se han utilizado sistemas *NAS*, utilizando para compartir los ficheros por la red el protocolo *Network File System (NFS)*.

Muchos hipervisores permiten trabajar con almacenamiento en red, como sistemas *SAN* utilizando *iSCSI* o *NAS* con *NFS*. De esta forma, utilizando varias máquinas con un hipervisor instalado, y un sistema de almacenamiento en red de alguno de estos tipos, es posible utilizar indistintamente una máquina virtual en cualquiera de las máquinas, permitiendo una alta disponibilidad en caso de fallo. De esta manera además se centralizan las tareas de respaldo de ficheros. Para los trabajos llevados a cabo en el desarrollo de esta tesis, se han utilizado sistemas *NAS* con *NFS* acoplados a dos servidores con el hipervisor *VMWare ESXi* instalado

3.6.4. GPGPU

La *GPGPU* o *General-Purpose Computing on Graphics Processing Units* es el concepto de aprovechar las capacidades de cómputo de las tarjetas gráficas para llevar a cabo otras tareas para las que originalmente estaban diseñadas. Como se ha hablado previamente en la Sección 1.1.1, gracias a *frameworks* como *CUDA* ha sido posible la utilización de las tarjetas gráficas para realizar cálculo científico, permitiendo aumentos de rendimiento considerables en la

ejecución de determinadas funciones de código.

Sin ocupar el tema principal de este trabajo, se realizó una colaboración en un trabajo que se sirve de este tipo de tecnologías, la aplicación *NMF-mGPU*[190]. En este trabajo se presenta una implementación del algoritmo *non-negative matrix factorization (NMF)*[152], un algoritmo iterativo de descomposición de matrices utilizado en bioinformática en análisis de agrupamiento de expresión génica o en minería de texto, entre otros, como se demostró con la herramienta *bioNMF*[224, 189]. Esta implementación ha sido llevada a cabo en tarjetas gráficas utilizando *CUDA*, con la posibilidad de utilizar simultáneamente multiples tarjetas a través del uso de paso de mensajes con *MPI*[267].

Capítulo 4

Conclusiones

El principal objetivo de la bioinformática es el de dotar a los biólogos de herramientas útiles para transformar los datos provenientes de experimentos en información útil que explique las características del mismo y permita generar nuevas hipótesis. En particular, en el ámbito de la genómica funcional, el objetivo es transformar datos masivos procedentes de las ómicas, previamente procesados y presentados, como por ejemplo listas de genes diferencialmente expresados, o proteínas presentes en la célula en un momento determinado, en información fácilmente asimilable acerca de los procesos que están teniendo lugar en la célula en ese instante y de esta forma ahondar en el conocimiento acerca de la función de los genes, las proteínas y sus interacciones. Estando englobado el presente trabajo en éste ámbito, se han abordado aspectos muy diferentes dentro del mismo, intentando finalmente crear una serie de herramientas, complementarias entre sí, que constituyan un marco de trabajo para la investigación con datos ómicos. Se han desarrollado diferentes métodos y herramientas bioinformáticas, comenzando por el procesamiento de datos crudos de un secuenciador alto rendimiento en experimentos de *RNA-Seq*, pudiendo generar a través de diferentes métodos de filtrado y comparación listas de genes que pueden ser estudiadas e interpretadas mediante las herramientas de comparación de perfiles, análisis funcional y estructural presentadas. Estos métodos de procesamiento de datos de secuenciadores también quieren servir de puente entre el estudio de la genómica y la proteómica. Además, se ha tenido en cuenta la importancia del análisis regulatorio de los genes a través del estudio de las interacciones entre ARN mensajero y micro ARNs, que puede condicionar en gran medida la expresión de genes extraídos de estos experimentos.

Las conclusiones finales de este trabajo son:

1. Se ha generado un flujo de trabajo automático para experimentos de *RNA-Seq* de diferentes repositorios públicos, que incluye alineamiento, cuantificación, filtrado de expresión de genes y búsqueda de mutaciones y nuevos sitios de *splicing* para generar bases de datos de aminoácidos útiles para experimentos de proteogenómica.

2. Se ha aplicado este flujo automático a experimentos *RNA-Seq* de diferentes repositorios públicos como *ENCODE* o *GEO*, y junto con resultados provenientes de otros tipos de experimentos de alto rendimiento (proteómica de *Shotgun*, *SRM* y *microarrays*) se han integrado en *dasHPPboard*, una aplicación web que permite visualizarlos y realizar búsquedas orientadas a proteogenómica.

3. Se ha desarrollado una nueva versión la aplicación web *GeneCodis*, una herramienta que permite el enriquecimiento funcional, modular y singular de listas de genes. Se han añadido nuevos organismos y bases de datos de anotaciones a la aplicación, así como nuevas funcionalidades como el análisis comparativo o la integración con la herramienta *GeneTerm Linker*. Se ha renovado la interfaz, añadiendo gráficos interactivos, nubes de términos y mejoras en las tablas de resultados. Se ha mejorado también el tiempo de respuesta de la aplicación gracias a la programación multihilo y al uso de la computación en *cluster*.

4. Se ha desarrollado un algoritmo de combinación de bases de datos de predicciones de interacciones entre ARN mensajeros y micro ARNs, que utiliza las puntuaciones normalizadas de estas bases de datos, las posiciones relativas de cada interacción dentro de las mismas y la proporción de interacciones validadas experimentalmente para generar un nuevo sistema de puntuaciones. Se ha desarrollado también una nueva metodología de comparación de este tipo de bases de datos.

5. A partir del desarrollo de este algoritmo, y con la inclusión de otro similar propuesto por colaboradores, se ha desarrollado una aplicación web para poder consultar información acerca de interacciones entre ARN mensajeros y micro ARNs, tanto de las puntuaciones de estos algoritmos como las de los utilizados a la hora de realizar las combinaciones, e información adicional sobre las interacciones validadas experimentalmente.

6. Se ha desarrollado una herramienta web de comparación de perfiles de expresión génica, que utiliza experimentos procesados de las bases de datos *GEO*, *Connectivity Map* y *DrugMatrix*, los asocia con diferente información acerca de fármacos, compuestos, enfermedades y expertos, y usando como entrada listas de genes procedentes de datos de

expresión génica, los experimentos de las diferentes bases de datos son ordenados por el parecido con esta última utilizando un algoritmo de reconocimiento de patrones, mostrando también la información relacionada con los mismos y de esta manera poder generar hipótesis en el contexto del reposicionamiento de fármacos.

7. Se ha creado una herramienta web que, a partir de información de asociación de estructuras a secuencias, y utilizando datos de anotaciones de diferentes bases de datos a nivel de secuencia, muestra a nivel visual estas anotaciones a nivel de estructura, gracias a un visor de moléculas *3D*, un panel de alineamiento entre secuencia y estructura, y un panel de anotaciones mostradas a nivel de secuencia.

4.1. Trabajo futuro

La información de bases de datos públicas incluída en las herramientas desarrolladas en este trabajo es la base de las mismas. Gran parte de estas bases de datos mantienen políticas de actualización frecuentes, por lo que resulta de gran importancia tener esto en cuenta. No todos los trabajos desarrollados han tenido el mismo éxito a la hora de llevar esto a cabo. En el caso de *3DBionotes*, a la hora de incluir una nueva base de datos de anotaciones se tiene que generar un nuevo *script* para procesar esa información correctamente. Una vez realizado, este *script* se incluye dentro del código de generación de bases de datos, que se ejecuta de forma periódica y reconstruye la base de datos interna de la herramienta. Este tipo de información es sencilla de procesar, a pesar de la variedad de formatos en los que se presenta, siendo el único problema el tiempo consumido en la reconstrucción de la misma, que aumenta con la inclusión de bases de datos de gran volumen. De esta forma, el proceso de actualización es automático y transparente para el mantenimiento de la herramienta. Además, las bases de datos que relacionan la información estructural y la de secuencia, así como la información almacenada de asociación de identificadores, son accedidas directamente, con la existencia únicamente de una caché para poder realizar accesos más rápidos a la información.

De forma similar, en la herramienta *NFFinder* la descarga de los experimentos se realiza a través de una *API* para *R* de *GEO*, llamada *GEOQuery*. Estos experimentos son posteriormente procesados también automáticamente, por lo que reconstruir la base de datos con firmas de *GEO* actualizadas sería sencillo. El único posible problema podría ser el de encontrar términos mal clasificados por la herramienta de procesamiento automático de texto, que tendrían que ser

encontradas e incluidas dentro de una lista de términos excluidos.

En el caso de la herramienta *dasHPPboard*, tanto el panel de la aplicación como la base de datos de búsqueda de genes y proteínas en experimentos, se basan en la información contenida en la estructura de carpetas de la aplicación. Para los datos de *RNA-Seq* del proyecto *ENCODE* la actualización se puede realizar de forma automática, con un *script* que descarga todos los ficheros de unas características determinadas, y realiza el flujo de análisis hasta generar los resultados que se introducen en esa estructura de carpetas de la aplicación. Con el resto de experimentos el proceso es más manual, ya que son datos muy heterogéneos. En estos caso, se propone como trabajo futuro automatizar al menos la inclusión de datos de algunas de estas fuentes. De especial interés sería la automatización del flujo de análisis de datos de proteómica de *Shotgun* procedentes de *ProteomeXchange*, lo que dotaría al portal de mucha más información proteómica con la que contrastar los datos transcriptómicos.

Las dos herramientas más problemáticas en este ámbito son *m^{3}RNA* y *GeneCodis*. Las bases de datos de interacciones de ARN mensajero con micro ARNs son muy heterogéneas, algunas requieren de registro, y no todas son descargables directamente para su procesamiento. Esto, junto con la diversidad de formatos de identificadores de genes y micro ARNs y los cambios de formato de tablas entre diferentes versiones de una misma base de datos, hacen de la actualización de los datos de esta herramienta un proceso obligatoriamente manual. En el caso de *GeneCodis*, el problema principal consiste en la gran cantidad de organismos y bases de datos distintas utilizadas en la herramienta. Gran cantidad de herramientas de enriquecimiento de listas de genes utilizan únicamente datos de *Gene Ontology* o *KEGG* por su facilidad de acceso y homogeneidad entre diferentes organismos. Actualmente el proceso de actualización en *GeneCodis* se tiene que realizar de una forma muy manual, debido a las particularidades de acceso y descarga de las bases de datos. En la actualidad existe un *script* para cada base de datos, pero luego los datos deben ensamblarse, chequeando posibles fallos o inconsistencias. La diversidad de identificadores utilizados en los nombres de genes se trata utilizando la herramienta *BioMart*[263], que permite unificar todos los datos. Se propone como trabajo futuro para estas herramientas explorar la posibilidad de utilizar servicios más modernos que *BioMart* para traducir identificadores de genes y obtener información de los mismos, tales como el servicio *MyGene.info*[312], que a través de un servicio web *RESTful* permite la obtención de datos en formato *JSON* de forma mucho más rápida y directa. También se propone abordar el problema de la heterogeneidad en las bases de datos, intentando generar *scripts* automáticos, o con una mínima intervención humana, que permitan realizar actualizaciones periódicas de los datos uti-

lizados por estas herramientas, así como almacenar y permitir utilizar diferentes versiones de las mismas.

Otro de los planes a corto plazo es el de generar una nueva visualización para el *backend* de *NFFinder*. En el momento del desarrollo de la herramienta, por encima de la capa de los servicios web de la herramienta, se generó una visualización web utilizando la herramienta *TIBCO Spotfire®*. Por problemas de licencias, en la actualidad los resultados de *NFFinder* se proporcionan directamente en la forma de salida *JSON*. Por este motivo se propone para el futuro una visualización realizada completamente en *HTML5* y *JavaScript* sin dependencias externas.

Aparte de estas mejoras en las herramientas ya existentes, los objetivos a medio plazo implican el desarrollo de nuevas herramientas dentro del ámbito de la genómica funcional. Destaca el interés por mejorar el refinado de resultados de enriquecimiento funcional, iniciado con la integración de la herramienta *GeneTerm Linker*, que mejora los resultados ofrecidos por *GeneCodis*, pero requiere de ajustes manuales para cada base de anotaciones utilizada, por lo que es complicado añadir nuevos recursos que permitan esta opción. Formas más sofisticadas de enriquecimiento, por ejemplo utilizando información acerca de la topología de rutas metabólicas como base, o buscando módulos dentro de grafos de relaciones entre anotaciones. Sería interesante también creación de nuevas herramientas que faciliten el análisis de experimentos de alto rendimiento en genómica, ya que gran cantidad de los avances en biología que se han producido en los últimos años utilizan estas técnicas, y pueden surgir nuevos tipos de flujos de análisis.

Bibliografía

[1] D. Abdulrehman y col. "YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface." En: *Nucleic acids research* 39.Database issue (2011), págs. D136-40.

[2] R. Aebersold, G. D. Bader, A. M. Edwards, J. E. Van Eyk, M. Kussmann, J. Qin y G. S. Omenn. "The biology/disease-driven human proteome project (B/D-HPP): Enabling protein research for the life sciences community". En: *Journal of Proteome Research* 12.1 (2013), págs. 23-27.

[3] R. Aebersold y M. Mann. "Mass spectrometry-based proteomics". En: *Nature* 422.6928 (2003), págs. 198-207.

[4] R. Agrawal, T. Imielinski y A. Swami. "Mining association rules between sets of items in large databases". En: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* 22.May (1993), págs. 207-216.

[5] F. Al-Shahrour, P. Minguez, J. Tárraga, I. Medina, E. Alloza, D. Montaner y J. Dopazo. "FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments." En: *Nucleic acids research* 35.Web Server issue (2007), W91-6.

[6] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts y P. Walter. *Molecular biology of the cell*. Garland Science, 2002.

[7] A. Alexa, J. Rahnenführer y T. Lengauer. "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure". En: *Bioinformatics* 22.13 (2006), págs. 1600-1607.

[8] R. B. Altman. "PharmGKB: a logical home for knowledge relating genotype to drug response phenotype". En: *Nature Genet.* 39.4 (2007), pág. 426.

[9] S. F. Altschul, W. Gish, W. T. Miller, E. W. Myers y D. J. Lipman. "Basic local alignment search tool". En: *J Mol Biol* 215.3 (1990), págs. 403-410.

[10] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott y A. Hamosh. "OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders". En: *Nucleic Acids Research* 43.D1 (2015), págs. D789-D798.

[11] V. Ambros y col. *A uniform system for microRNA annotation.* 2003.

[12] S. Anders, P. T. Pyl y W. Huber. "HTSeq-A Python framework to work with high-throughput sequencing data". En: *Bioinformatics* 31.2 (2015), págs. 166-169.

[13] S. Andrews. *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. 2015.

[14] A. R. Aronson. "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." En: *Proceedings. AMIA Symposium* (2001), págs. 17-21.

[15] A. R. Aronson y F.-M. Lang. "An overview of MetaMap: historical perspective and recent advances." En: *Journal of the American Medical Informatics Association : JAMIA* 17.3 (2010), págs. 229-36.

[16] M. Ashburner y col. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." En: *Nature genetics* 25.1 (2000), págs. 25-29.

[17] A Bairoch y R Apweiler. "The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000". En: *Nucleic acids research* 28.1 (2000), págs. 45-48.

[18] H. van Bakel, C. Nislow, B. J. Blencowe y T. R. Hughes. "Most "dark matter"transcripts are associated with known genes". En: *PLoS Biology* 8.5 (2010). Ed. por S. R. Eddy, e1000371.

[19] M. Baker. "Next-generation sequencing: adjusting to data overload." En: *Nature methods* 7.7 (2010), págs. 495-499.

[20] B. E. Bammes, R. H. Rochat, J. Jakana, D. H. Chen y W. Chiu. "Direct electron detection yields cryo-EM reconstructions at resolutions beyond 3/4 Nyquist frequency". En: *Journal of Structural Biology* 177.3 (2012), págs. 589-601.

[21] T. A. Ban. "The role of serendipity in drug discovery". En: *Dialogues in Clinical Neuroscience* 8.3 (2006), págs. 335-344.

[22] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt y A. Warfield. "Xen and the art of virtualization". En: *ACM SIGOPS Operating Systems Review* 37.5 (2003), pág. 164.

[23] J. Barretina y col. "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity." En: *Nature* 483.7391 (2012), págs. 603-7.

[24] T. Barrett y col. "NCBI GEO: archive for functional genomics data sets–update." En: *Nucleic acids research* 41.Database issue (2013), págs. D991-5.

[25] D. P. Bartel. "MicroRNAs: genomics, biogenesis, mechanism, and function." En: *Cell* 116.2 (2004), págs. 281-97.

[26] S. Bauer, S. Grossmann, M. Vingron y P. N. Robinson. "Ontologizer 2.0 - A multifunctional tool for GO term enrichment analysis and data exploration". En: *Bioinformatics* 24.14 (2008), págs. 1650-1651.

[27] Y. Benjamini e Y. Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". En: *Journal of the royal statistical society. Series B (Methodological)* (1995), págs. 289-300.

[28] D. R. Bentley y col. "Accurate whole human genome sequencing using reversible terminator chemistry." En: *Nature* 456.7218 (2008), págs. 53-9.

[29] H. Berman, K. Henrick y H. Nakamura. "Announcing the worldwide Protein Data Bank." En: *Nature Structural Biology* 10.12 (2003), pág. 980.

[30] H. M. Berman, G. J. Kleywegt, H. Nakamura y J. L. Markley. "The Protein Data Bank archive as an open data resource". En: *Journal of Computer-Aided Molecular Design* 28.10 (2014), págs. 1009-1014.

[31] D. Betel, M. Wilson, A. Gabow, D. S. Marks y C. Sander. "The microRNA.org resource: Targets and expression". En: *Nucleic Acids Research* 36.SUPPL. 1 (2008), págs. D149-D153.

[32] B. Bioinformatics. "FASTQC: A quality control tool for high throughput sequence data". En: *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/* (2012).

[33] D. L. Black. "Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology". En: *Cell* 103.3 (2000), págs. 367-370.

[34] J. A. Blake y col. "Gene ontology consortium: Going forward". En: *Nucleic Acids Research* 43.D1 (2015), págs. D1049-D1056.

[35] O. Bodenreider. "The Unified Medical Language System (UMLS): integrating biomedical terminology". En: *Nucleic Acids Research* 32.suppl 1 (2004), págs. D267-D270.

[36] C. Borgelt, X. Yang, R. Nogales-Cadenas, P. Carmona-Saez y A. Pascual-Montano. "Finding closed frequent item sets by intersecting transactions". En: *Proceedings of the 14th International Conference on Extending Database Technology - EDBT/ICDT '11*. New York, New York, USA: ACM Press, 2011, pág. 367.

[37] A Brazma y col. "Minimum information about a microarray experiment (MIAME)-toward standards for microarray data". En: *Nat Genet* 29.december (2001), págs. 365-371.

[38] J. Brennecke, A. Stark, R. B. Russell y S. M. Cohen. "Principles of microRNA-target recognition." En: *PLoS biology* 3.3 (2005), e85.

[39] Broad Institute. *Picard tools*. 2016.

[40] P. Brodersen y O. Voinnet. "Revisiting the principles of microRNA target recognition and mode of action." En: *Nature reviews. Molecular cell biology* 10.2 (2009), págs. 141-8.

[41] O. P. Brown. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray". En: *Science* 270.5235 (1995), págs. 467-470.

[42] R. Bumgarner. "Overview of dna microarrays: Types, applications, and their future". En: *Current Protocols in Molecular Biology* Chapter 22.SUPPL.101 (2013), Unit 22.1.

[43] P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo y A. Pascual-Montano. "GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists." En: *Genome biology* 8.1 (2007), R3.

[44] M. J. Chaisson, D. Brinza y P. A. Pevzner. "De novo fragment assembly with short mate-paired reads: Does the read length matter?" En: *Genome Research* 19.2 (2009), págs. 336-346.

[45] D. Chen, W. Guo, Z. Qiu, Q. Wang, Y. Li, L. Liang, L. Liu, S. Huang, Y. Zhao y X. He. "MicroRNA-30d-5p inhibits tumour cell proliferation and motility by directly targeting CCNE2 in non-small cell lung cancer". En: *Cancer letters* 362.2 (2015), págs. 208-217.

[46] A. M. Cheng, M. W. Byrom, J. Shelton y L. P. Ford. "Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis." En: *Nucleic acids research* 33.4 (2005), págs. 1290-7.

[47] S. W. Chi, J. B. Zang, A. Mele y R. B. Darnell. "Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps." En: *Nature* 460.7254 (2009), págs. 479-86.

[48] C. H. Chou y col. "miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database". En: *Nucleic Acids Research* 44.D1 (2016), págs. D239-D247.

[49] G. A. Churchill. "Using ANOVA to analyze microarray data". En: *BioTechniques* 37.2 (2004), págs. 173-177.

[50] P. Cingolani, A. Platts, L. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu y D. M. Ruden. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3". En: *Fly* 6.2 (2012), págs. 80-92.

[51] *CLUE Platform [clue.io]*. `https://clue.io/clue`. (Accessed on 03/24/2017).

[52] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer y P. M. Rice. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". En: *Nucleic Acids Research* 38.6 (2009), págs. 1767-1771.

[53] C. Coronnello y P. V. Benos. "ComiR: combinatorial microRNA target prediction tool". En: *Nucleic Acids Research* 4 (2013), págs. 1-6.

[54] R. Craig y R. C. Beavis. "TANDEM: Matching proteins with tandem mass spectra". En: *Bioinformatics* 20.9 (2004), págs. 1466-1467.

[55] M. Dai y col. "NGSQC: cross-platform quality analysis pipeline for deep sequencing data". En: *BMC Genomics* 11.Suppl 4 (2010), S7.

[56] J. Daily. "Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments". En: *BMC Bioinformatics* 17 (81) (2016).

[57] J. Davis y M. Goadrich. "The relationship between Precision-Recall and ROC curves". En: *Proceedings of the 23rd international conference on Machine learning ICML 06*. ICML '06 10.2 (2006), págs. 233-240.

[58] S. Davis y P. S. Meltzer. “GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor”. En: *Bioinformatics* 23 (2007), págs. 1846-1847.

[59] J. M. De la Rosa-Trevín, J. Otón, R. Marabini, A. Zaldívar, J. Vargas, J. M. Carazo y C. O. S. Sorzano. “Xmipp 3.0: An improved software suite for image processing in electron microscopy”. En: *Journal of Structural Biology* 184.2 (2013), págs. 321-328.

[60] R. P. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen y R. Etzioni. “Combining results of microarray experiments: a rank aggregation approach.” En: *Statistical applications in genetics and molecular biology* 5 (2006), Article15.

[61] F. Desiere, E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich y R. Aebersold. “The PeptideAtlas project.” En: *Nucleic acids research* 34.Database issue (2006), págs. D655-8.

[62] F. E. Dewey y col. “Clinical Interpretation and Implications of Whole-Genome Sequencing”. En: *JAMA* 311.10 (2014), pág. 1035.

[63] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson y T. R. Gingeras. “STAR: Ultrafast universal RNA-seq aligner”. En: *Bioinformatics* 29.1 (2013), págs. 15-21.

[64] J. G. Doench y P. a. Sharp. “Specificity of microRNA target selection in translational repression.” En: *Genes & development* 18.5 (2004), págs. 504-11.

[65] K. Dreij, K. Rhrissorrakrai, K. C. Gunsalus, N. E. Geacintov y D. A. Scicchitano. “Benzo [a] pyrene diol epoxide stimulates an inflammatory response in normal human lung fibroblasts through a p53 and JNK mediated pathway”. En: *Carcinogenesis* (2010), bgq073.

[66] D. Dressman, H. Yan, G. Traverso, K. W. Kinzler y B. Vogelstein. “Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations.” En: *Proceedings of the National Academy of Sciences of the United States of America* 100.15 (2003), págs. 8817-8822.

[67] J. T. Dudley, T. Deshpande y A. J. Butte. “Exploiting drug-disease relationships for computational drug repositioning.” En: *Briefings in bioinformatics* 12.4 (2011), págs. 303-11.

[68] H. Dweep, C. Sticht, P. Pandey y N. Gretz. “miRWalk–database: prediction of possible miRNA binding sites by ”walking”the genes of three genomes.” En: *Journal of biomedical informatics* 44.5 (2011), págs. 839-47.

[69] J. Eid y col. “Real-Time DNA Sequencing from Single Polymerase Molecules”. En: *Science* 323.5910 (2009), págs. 133-138.

[70] M. B. Eisen, P. T. Spellman, P. O. Brown y D Botstein. “Cluster analysis and display of genome-wide expression patterns”. En: *Proc Natl Acad Sci USA* 95.25 (1998), págs. 14863-14868.

[71] S. Ekins, A. J. Williams, M. D. Krasowski y J. S. Freundlich. “In silico repositioning of approved drugs for rare and neglected diseases.” En: *Drug discovery today* 16.7-8 (2011), págs. 298-310.

[72] J. E. Elias y S. P. Gygi. "Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry". En: *Nature Methods* 4.3 (2007), págs. 207-214.

[73] J. K. Eng, A. L. McCormack y J. R. Yates. "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database". En: *Journal of the American Society for Mass Spectrometry* 5.11 (1994), págs. 976-989.

[74] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander y D. S. Marks. "MicroRNA targets in Drosophila." En: *Genome biology* 5.1 (2003), R1.

[75] B. Ewing, B. Ewing, L. Hillier, L. Hillier, M. C. Wendl, M. C. Wendl, P. Green y P. Green. "Base-Calling of Automated Sequencer Traces Using". En: *Genome Research* 8.206 (2005), págs. 175-185.

[76] B. Ewing y P. Green. "Base-calling of automated sequencer traces using phred. II. Error probabilities". En: *Genome Research* 8.3 (1998), págs. 186-194.

[77] M. Fedurco, A. Romieu, S. Williams, I. Lawrence y G. Turcatti. "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies." En: *Nucleic acids research* 34.3 (2006), e22.

[78] J. Fenn, M. Mann, C. Meng, F. Shek y C. Whitehouse. "Electrospray ionization for mass spectrometry of large biomolecules". En: *Science* 246.6 (1989), págs. 64-71.

[79] *FGED: MINSEQE*. `http://www.fged.org/projects/minseqe/`. (Accessed on 03/24/2017).

[80] R. D. Finn y col. "InterPro in 2017-beyond protein family and domain annotations." En: *Nucleic acids research* 45.D1 (2016), gkw1107.

[81] C. Fontanillo, R. Nogales-Cadenas, A. Pascual-Montano y J. De Las Rivas. "Functional Analysis beyond Enrichment: Non-Redundant Reciprocal Linkage of Genes and Biological Terms". En: *PLoS ONE* 6.9 (2011). Ed. por D. Bhattacharya, e24289.

[82] M. Fowler. *Patterns of enterprise application architecture*. Addison-Wesley, 2009, pág. 533.

[83] R. C. Friedman, K. K.-H. Farh, C. B. Burge y D. P. Bartel. "Most mammalian mRNAs are conserved targets of microRNAs." En: *Genome research* 19.1 (2009), págs. 92-105.

[84] M. J. Fullwood, C. L. Wei, E. T. Liu e Y. Ruan. *Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses*. 2009.

[85] D. Gaidatzis, E. van Nimwegen, J. Hausser y M. Zavolan. "Inference of miRNA targets using evolutionary conservation and pathway analysis." En: *BMC bioinformatics* 8 (2007), pág. 69.

[86] E. R. Gamazon, H.-K. Im, S. Duan, Y. A. Lussier, N. J. Cox, M. E. Dolan y W. Zhang. "Exprtarget: an integrative approach to predicting human microRNA targets." En: *PloS one* 5.10 (2010), e13534.

[87] P. Gaudet y col. "neXtProt: organizing protein knowledge in the context of human proteome projects." En: *Journal of proteome research* 12.1 (2013), págs. 293-8.

[88] L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi y S. H. Bryant. "Open mass spectrometry search algorithm". En: *Journal of Proteome Research* 3.5 (2004), págs. 958-964.

[89] T. Geiger, J. Cox y M. Mann. "Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation". En: *Mol Cell Proteomics* 9.10 (2010), págs. 2252-2261.

[90] R. C. Gentleman y col. "Bioconductor: open software development for computational biology and bioinformatics." En: *Genome biology* 5.10 (2004), R80.

[91] *GEO Data - Help - ArrayExpress - EMBL-EBI*. `https://www.ebi.ac.uk/arrayexpress/help/GEO_data.html`. (Accessed on 03/24/2017).

[92] T. C. Glenn. "Field guide to next-generation DNA sequencers". En: *Molecular Ecology Resources* 11.5 (2011), págs. 759-769.

[93] G. H. Golub y C. F. van Loan. *Matrix Computations*. Johns Hopkins University Press, 1989, pág. 694.

[94] J. Gomez y R. Jimenez. "Sequence, a BioJS component for visualising sequences". En: *F1000Research* 3 (2014), pág. 52.

[95] J. Gómez y col. "BioJS: An open source JavaScript framework for biological data visualization". En: *Bioinformatics* 29.8 (2013), págs. 1103-1104.

[96] R. J. A. Goode, S. Yu, A. Kannan, J. H. Christiansen, A. Beitz, W. S. Hancock, E. Nice y A. I. Smith. "The proteome browser web portal." En: *Journal of proteome research* 12.1 (2013), págs. 172-8.

[97] S. Griffiths-Jones, H. K. Saini, S. van Dongen y A. J. Enright. "miRBase: tools for microRNA genomics." En: *Nucleic acids research* 36.Database issue (2008), págs. D154-8.

[98] S Gröschel y col. "An oncogenic enhancer-rearrangement causes concomitant deregulation of EVI1 and GATA2 in leukemia". En: *Cell* In Press.2 (2014), págs. 369-381.

[99] J. Guinney y col. "The consensus molecular subtypes of colorectal cancer". En: *Nature Medicine* 21.11 (2015), págs. 1350-1356.

[100] F. Guo y col. "CAPER: a chromosome-assembled human proteome browsER." En: *Journal of proteome research* 12.1 (2013), págs. 179-86.

[101] E. Guruceaga, M. Sanchez Del Pino, F. J. Corrales y V. Segura. "Prediction of a missing protein expression map in the context of the Human Proteome Project." En: *Journal of proteome research* (2015).

[102] T. Hansen y col. "Brain expressed microRNAs implicated in schizophrenia etiology." En: *PloS one* 2.9 (2007), e873.

[103] R. M. Hanson, J. Prilusky, Z. Renjian, T. Nakane y J. L. Sussman. *JSmol and the next-generation web-based representation of 3D molecular structure as applied to proteopedia*. 2013.

[104] T. D. Harris y col. "Single-molecule DNA sequencing of a viral genome." En: *Science (New York, N.Y.)* 320.5872 (2008), págs. 106-9.

[105] J. Harrow y col. “GENCODE: the reference human genome annotation for The ENCODE Project.” En: *Genome research* 22.9 (2012), págs. 1760-74.

[106] A. Hassibi, C. Contag, M. O. Vlad, M. Hafezi, T. H. Lee, R. W. Davis y N. Pourmand. “Bioluminescence regenerative cycle (BRC) system: theoretical considerations for nucleic acid quantification assays.” En: *Biophysical chemistry* 116.3 (2005), págs. 175-85.

[107] R. Haw, H. Hermjakob, P. D'Eustachio y L. Stein. “Reactome pathway analysis to enrich biological discovery in proteomics data sets”. En: *Proteomics* 11.18 (2011), págs. 3598-3613.

[108] E. C. Hayden. “The $1,000 genome”. En: *Nature* 507 (2014), pág. 295.

[109] D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. V. Oudenaarden, A. van Oudenaarden y S. a. Teichmann. “RNA sequencing reveals two major classes of gene expression levels in metazoan cells”. En: *Molecular Systems Biology* 7.497 (2011), págs. 1-9.

[110] M. Helsley. “LXC: Linux container tools”. En: *IBM devloperWorks Technical Library* (2009), págs. 1-10.

[111] M. Hewett, D. E. Oliver, D. L. Rubin, K. L. Easton, J. M. Stuart, R. B. Altman y T. E. Klein. “PharmGKB: The pharmacogenetics knowledge base”. En: *Nucleic Acids Research* 30.1 (2002), págs. 163-165.

[112] P. Hieter y M. Boguski. “Functional Genomics: It’s All How You Read It”. En: *Science* 278.5338 (1997), págs. 601-602.

[113] I. L. Hofacker. “Vienna RNA secondary structure server”. En: *Nucleic Acids Research* 31.13 (2003), págs. 3429-3431.

[114] E. de. Hoffmann y V. Stroobant. *Mass spectrometry : principles and applications*. J. Wiley, 2007, pág. 489.

[115] P. Hogeweg. “The roots of bioinformatics in theoretical biology”. En: *PLoS Computational Biology* 7.3 (2011). Ed. por D. B. Searls, e1002021.

[116] J. D. Hoheisel. “Microarray technology: beyond transcript profiling and genotype analysis.” En: *Nature reviews. Genetics* 7.3 (2006), págs. 200-210.

[117] P. V. Hornbeck, B. Zhang, B. Murray, J. M. Kornhauser, V. Latham y E. Skrzypek. “PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations”. En: *Nucleic Acids Research* 43.D1 (2015), págs. D512-D520.

[118] S.-D. Hsu y col. “miRTarBase: a database curates experimentally validated microRNA-target interactions.” En: *Nucleic acids research* 39.Database issue (2011), págs. D163-9.

[119] S. D. Hsu y col. “MiRTarBase update 2014: An information resource for experimentally validated miRNA-target interactions”. En: *Nucleic Acids Research* 42.D1 (2014), págs. D78-85.

[120] D. W. Huang, B. T. Sherman y R. A. Lempicki. “Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.” En: *Nucleic acids research* 37.1 (2009), págs. 1-13.

[121] D. W. Huang, B. T. Sherman y R. A. Lempicki. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." En: *Nature protocols* 4.1 (2009), págs. 44-57.

[122] Q. Huang y col. "The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis." En: *Nature cell biology* 10.2 (2008), págs. 202-10.

[123] S. Huang, J. Zhang, R. Li, W. Zhang, Z. He, T. W. Lam, Z. Peng y S. M. Yiu. "SOAPsplice: Genome-wide ab initio detection of splice junctions from RNA-Seq data". En: *Frontiers in Genetics* 2.JULY (2011), pág. 46.

[124] Hubbard, T. et al. "The Ensembl genome database project." En: *Nucleic acids research* 30.1 (2002), págs. 38-41.

[125] T. R. Hughes y col. "Functional Discovery via a Compendium of Expression Profiles". En: *Cell* 102.1 (2000), págs. 109-126.

[126] J. D. Jaffe, H. C. Berg y G. M. Church. "Proteogenomic mapping as a complementary method to perform genome annotation". En: *Proteomics* 4.1 (2004), págs. 59-77.

[127] M. Janitz. *Next-generation genome sequencing : towards personalized medicine*. Wiley-VCH, 2008, pág. 260.

[128] P. a. Jaskowiak, R. J.G. B. Campello e I. G. Costa. "On the selection of appropriate distances for gene expression data clustering." En: *BMC bioinformatics* 15 Suppl 2.Suppl 2 (2014), S2.

[129] S.-K. Jeong y col. "GenomewidePDB, a proteomic database exploring the comprehensive protein parts list and transcriptome landscape in human chromosomes." En: *Journal of proteome research* 12.1 (2013), págs. 106-11.

[130] I. T. Jolliffe. *Principal component analysis*. Springer, 1986, pág. 487.

[131] J. Kappel, A Velte y T Velte. *Microsoft Virtualization with Hyper-V*. McGraw Hill, 2009, pág. 448.

[132] M. Karas y F. Hillenkamp. "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons." En: *Analytical chemistry* 60.20 (1988), págs. 2299-2301.

[133] N. L. Kelleher. "Peer Reviewed: Top-Down Proteomics". En: *Analytical Chemistry* 76.11 (2004), 196 A-203 A.

[134] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul y E. Segal. "The role of site accessibility in microRNA target recognition." En: *Nature genetics* 39.10 (2007), págs. 1278-1284.

[135] P. Khatri, S. Draghici, G. C. Ostermeier y S. a. Krawetz. "Profiling gene expression using onto-express." En: *Genomics* 79.2 (2002), págs. 266-70.

[136] H Kiaris, D. Spandidos, A. Jones, E. Vaughan y J. Field. "Mutations, expression and genomic instability of the H-ras proto-oncogene in squamous cell carcinomas of the head and neck". En: *British journal of cancer* 72.1 (1995), págs. 123-128.

[137] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley y S. L. Salzberg. "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." En: *Genome biology* 14.4 (2013), R36.

[138] M.-S. Kim y col. "A draft map of the human proteome." En: *Nature* 509.7502 (2014), págs. 575-81.

[139] A. Kivity, Y. Kamay, D. Laor, U. Lublin y A. Liguori. "kvm: the Linux virtual machine monitor". En: 1 (2007), págs. 225-230.

[140] R. Kodzius y col. "CAGE: cap analysis of gene expression." En: *Nature methods* 3.3 (2006), págs. 211-222.

[141] N. Kolesnikov y col. "ArrayExpress update-simplifying data submissions." En: *Nucleic acids research* 43.Database issue (2015), págs. D1113-6.

[142] I. Kononenko. "Machine Learning: ECML-94". En: *Machine Learning: ECML-94* 784 (1994), págs. 171-182.

[143] A. Kozomara y S. Griffiths-Jones. "miRBase: annotating high confidence microRNAs using deep sequencing data." En: *Nucleic acids research* 42.Database issue (2014), págs. D68-73.

[144] A. Kozomara y S. Griffiths-Jones. "miRBase: integrating microRNA annotation and deep-sequencing data." En: *Nucleic acids research* 39.Database issue (2011), págs. D152-7.

[145] G. E. Krasner y S. T. Pope. "A Cookbook for Using the Model- View-Controller User Interface Paradigm in Smalltalk-80". En: *Joop Journal Of Object Oriented Programming* 1.3 (1988), págs. 26-49.

[146] W. Kühlbrandt. "The Resolution Revolution". En: *Science* 343.March (2014), págs. 1443-1444.

[147] H. Lab y Hannon Lab. *FASTX-Toolkit*. `http://hannonlab.cshl.edu/fastx{_}toolkit/index.html`. (Accessed on 03/24/2017).

[148] A. Laganà, S. Forte, F. Russo, R. Giugno, A. Pulvirenti y A. Ferro. "Prediction of human targets for viral-encoded microRNAs by thermodynamics and empirical constraints." En: *Journal of RNAi and gene silencing : an international journal of RNA and gene targeting research* 6.1 (2010), págs. 379-385.

[149] J. Lamb y col. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease." En: *Science (New York, N.Y.)* 313.5795 (2006), págs. 1929-35.

[150] E. S. Lander y col. "Initial sequencing and analysis of the human genome". En: *Nature (London)* 409.6822 (2001), págs. 860-921.

[151] C. L. Lawson y col. "EMDataBank unified data resource for 3DEM". En: *Nucleic Acids Research* 44.D1 (2016), págs. D396-D403.

[152] D. D. Lee y H. S. Seung. "Learning the parts of objects by non-negative matrix factorization." En: *Nature* 401.6755 (1999), págs. 788-91.

[153] J.-H. Lee, D. G. Kim, T. J. Bae, K. Rho, J.-T. Kim, J.-J. Lee, Y. Jang, B. C. Kim, K. M. Park y S. Kim. "CDA: combinatorial drug discovery using transcriptional response modules." En: *PloS one* 7.8 (2012), e42573.

[154] R. C. Lee, R. L. Feinbaum y V Ambros. "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." En: *Cell* 75.5 (1993), págs. 843-54.

[155] P. Legrain y col. "The human proteome project: current state and future direction." En: *Molecular & cellular proteomics : MCP* 10.7 (2011), pág. M111.009993.

[156] R. Leinonen, H. Sugawara y M. Shumway. "The sequence read archive." En: *Nucleic acids research* 39.Database issue (2011), págs. D19-21.

[157] J. Z. Levin, M. Yassour, X. Adiconis, C. Nusbaum, D. A. Thompson, N. Friedman, A. Gnirke y A. Regev. "Comprehensive comparative analysis of strand-specific RNA sequencing methods." En: *Nature methods* 7.9 (2010), págs. 709-15.

[158] B. P. Lewis, C. B. Burge y D. P. Bartel. "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." En: *Cell* 120.1 (2005), págs. 15-20.

[159] B. P. Lewis, I.-h. Shih, M. W. Jones-Rhoades, D. P. Bartel y C. B. Burge. "Prediction of mammalian microRNA targets." En: *Cell* 115.7 (2003), págs. 787-98.

[160] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis y R. Durbin. "The Sequence Alignment/Map format and SAMtools." En: *Bioinformatics (Oxford, England)* 25.16 (2009), págs. 2078-9.

[161] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis y R. Durbin. "The Sequence Alignment/Map format and SAMtools". En: *Bioinformatics* 25.16 (2009), págs. 2078-2079.

[162] J. H. Li, S. Liu, H. Zhou, L. H. Qu y J. H. Yang. "StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data". En: *Nucleic Acids Research* 42.D1 (2014), págs. D92-7.

[163] Y. Liao, G. K. Smyth y W. Shi. "FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features". En: *Bioinformatics* 30.7 (2014), págs. 923-930.

[164] R. F. Ling y J. W. Pratt. "The Accuracy of Peizer Approximations to the Hypergeometric Distribution, with Comparisons to Some other Approximations". En: *Journal of the American Statistical Association* 79.385 (1984), págs. 49-60.

[165] Y. Liu, D. L. Maskell y B. Schmidt. "CUDASW++: optimizing Smith-Waterman sequence database searches for CUDA-enabled graphics processing units". En: *BMC Research Notes* 2:73 (2009).

[166] Y. Liu, A. Wirawan y B. Schmidt. "CUDASW++ 3.0: accelerating Smith-Waterman protein database search by coupling CPU and GPU SIMD instructions". En: *BMC Bioinformatics* 14:117 (2013).

[167] Y. Liu y B. Schmidt. "SWAPHI: Smith-Waterman protein database search on Xeon Phi coprocessors". En: *25th IEEE International Conference on Application-specific Systems, Architectures and Processors (ASAP 2014)*. 2014.

[168] Y. Liu, B. Schmidt y D. L. Maskell. "CUDASW++2.0: enhanced Smith-Waterman protein database search on CUDA-enabled GPUs based on SIMT and virtualized SIMD abstractions". En: *BMC Research Notes* 3.1 (2010), págs. 1-12.

[169] M. I. Love, W. Huber y S. Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." En: *Genome biology* 15.12 (2014), pág. 550.

[170] B. Ma y R. Johnson. "De Novo Sequencing and Homology Searching". En: *Molecular & Cellular Proteomics* 11.2 (2012), O111.014902-O111.014902.

[171] J. Ma, X. Guo, S. Zhang, H. Liu, J. Lu, Z. Dong, K. Liu y L. Ming. "Trichostatin A, a histone deacetylase inhibitor, suppresses proliferation and promotes apoptosis of esophageal squamous cell lines". En: *Molecular medicine reports* 11.6 (2015), págs. 4525-4531.

[172] J. P. de Magalhães, J. Curado y G. M. Church. "Meta-analysis of age-related gene expression profiles identifies common signatures of aging". En: *Bioinformatics* 25.7 (2009), págs. 875-881.

[173] D. Maglott, J. Ostell, K. D. Pruitt y T. Tatusova. "Entrez gene: Gene-centered information at NCBI". En: *Nucleic Acids Research* 39.SUPPL. 1 (2011), págs. D26-31.

[174] M. Magrane y U. Consortium. "UniProt Knowledgebase: a hub of integrated protein data." En: *Database : the journal of biological databases and curation* 2011.0 (2011), bar009.

[175] M Maragkakis y col. "DIANA-microT web server: elucidating microRNA functions through target prediction." En: *Nucleic acids research* 37.Web Server issue (2009), W273-6.

[176] E. R. Mardis. "Anticipating the 1,000 dollar genome." En: *Genome biology* 7.7 (2006), pág. 112.

[177] M. Margulies y col. "Genome sequencing in microfabricated high-density picolitre reactors." En: *Nature* 437.7057 (2005), págs. 376-80.

[178] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens e Y. Gilad. "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays". En: *Genome Research* 18.9 (2008), págs. 1509-1517.

[179] L. Martens y col. "mzML–a Community Standard for Mass Spectrometry Data". En: *Molecular & Cellular Proteomics* 10.1 (2011), R110.000133-R110.000133.

[180] V. Marx. "Biology: The big challenges of big data". En: *Nature* 498.7453 (2013), págs. 255-260.

[181] V. Matys. "TRANSFAC(R): transcriptional regulation, from patterns to profiles". En: *Nucleic Acids Research* 31.1 (2003), págs. 374-378.

[182] M. N. McCall, B. M. Bolstad y R. A. Irizarry. "Frozen robust multiarray analysis (fRMA)." En: *Biostatistics (Oxford, England)* 11.2 (2010), págs. 242-53.

[183] M. N. McCall y R. A. Irizarry. "Consolidated strategy for the analysis of microarray spike-in data." En: *Nucleic acids research* 36.17 (2008), e108.

[184] M. N. McCall, H. A. Jaffee, S. J. Zelisko, N. Sinha, G. Hooiveld, R. A. Irizarry y M. J. Zilliox. "The Gene Expression Barcode 3.0: improved data processing and mining tools." En: *Nucleic acids research* 42.Database issue (2014), págs. D938-43.

[185] D. J. McCarthy, Y. Chen y G. K. Smyth. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". En: *Nucleic Acids Research* 40.10 (2012), págs. 4288-4297.

[186] A. McKenna y col. "The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data". En: *Genome Research* 20.9 (2010), págs. 1297-1303.

[187] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek y F. Cunningham. "The Ensembl Variant Effect Predictor". En: *bioRxiv* 17.1 (2016), pág. 042374.

[188] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek y F. Cunningham. "Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor." En: *Bioinformatics (Oxford, England)* 26.16 (2010), págs. 2069-70.

[189] E. Mejía-Roa, P. Carmona-Saez, R. Nogales, C. Vicente, M. Vázquez, X. Y. Yang, C. García, F. Tirado y A. Pascual-Montano. "bioNMF: a web-based tool for nonnegative matrix factorization in biology." En: *Nucleic acids research* 36.Web Server issue (2008), W523-8.

[190] E. Mejía-roa, D. Tabas-Madrid, J. Setoain, C. García, F. Tirado y A. Pascual-Montano. "NMF-mGPU : non-negative matrix factorization on multi-GPU systems". En: *BMC Bioinformatics* 16.1 (2015), págs. 1-12.

[191] B. H. M. Meldal y col. "The complex portal - An encyclopaedia of macromolecular complexes". En: *Nucleic Acids Research* 43.D1 (2015), págs. D479-D484.

[192] M. R. Mendoza, G. C. da Fonseca, G. Loss-Morais, R. Alves, R. Margis y A. L. C. Bazzan. "RFMirTarget: Predicting Human MicroRNA Target Genes with a Random Forest Classifier". En: *PLoS ONE* 8.7 (2013), e70153.

[193] D Merkel. "Docker: lightweight linux containers for consistent development and deployment". En: *Linux Journal* (2014).

[194] B. Merriman, I. T. R&D Team y J. M. Rothberg. "Progress in Ion Torrent semiconductor chip based sequencing". En: *Electrophoresis* 33.23 (2012), págs. 3397-3417.

[195] F. Mertes, A. ElSharawy, S. Sauer, J. M.L. M. van Helvoort, P. J. van der Zaag, A. Franke, M. Nilsson, H. Lehrach y A. J. Brookes. *Targeted enrichment of genomic DNA regions for next-generation sequencing*. 2011.

[196] M. L. Metzker. “Emerging technologies in DNA sequencing”. En: *Genome Research* 15.12 (2005), págs. 1767-1776.

[197] H. Mi, S. Poudel, A. Muruganujan, J. T. Casagrande y P. D. Thomas. “PANTHER version 10: Expanded protein families and functions, and analysis tools”. En: *Nucleic Acids Research* 44.D1 (2016), págs. D336-D342.

[198] R. Mitra y S. Bandyopadhyay. “MultiMiTar: A novel multi objective optimization based miRNA-target prediction method”. En: *PLoS ONE* 6.9 (2011), e24583.

[199] B. Modrek y C. Lee. “A genomic view of alternative splicing.” En: *Nature genetics* 30.1 (2002), págs. 13-19.

[200] G. E. Moore. “Cramming more components onto integrated circuits”. En: *Electronics* 38.8 (1965).

[201] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer y B. Wold. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. En: *Nature Methods* 5.7 (2008), págs. 621-628.

[202] R. Mosca, J. Tenorio-Laranga, R. Olivella, V. Alcalde, A. Céol, M. Soler-López y P. Aloy. “dSysMap: exploring the edgetic role of disease mutations”. En: *Nature Methods* 12.3 (2015), págs. 167-168.

[203] P. Muir y col. “The real cost of sequencing: scaling computation to keep pace with data generation.” En: *Genome biology* 17.1 (2016), pág. 53.

[204] A. Muller y S. Wilson. *Virtualization with vmware ESX server*. Syngress Publishing, 2005, pág. 608.

[205] A. Muniategui, R. Nogales-Cadenas, M. Vázquez, X. L Aranguren, X. Agirre, A. Luttun, F. Prosper, A. Pascual-Montano y A. Rubio. “Quantification of miRNA-mRNA interactions.” En: *PLoS ONE* 7.2 (2012). Ed. por P. Provero, e30766.

[206] R. M. Myers y col. “A user’s guide to the Encyclopedia of DNA elements (ENCODE)”. En: *PLoS Biology* 9.4 (2011).

[207] S. H. Nagaraj, N. Waddell, A. K. Madugundu, S. Wood, A. Jones, R. A. Mandyam, K. Nones, J. V. Pearson y S. M. Grimmond. “PGTools: A Software Suite for Proteogenomic Data Analysis and Visualization.” En: *Journal of proteome research* 14.5 (2015), págs. 2255-66.

[208] V. Narasimhan, P. Danecek, A. Scally, Y. Xue, C. Tyler-Smith y R. Durbin. “BCFtools / RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data”. En: *Bioinformatics* 32.11 (2016), págs. 1749-1751.

[209] P. Navarro y J. Vazquez. “A refined method to calculate false discovery rates for peptide identification using decoy databases”. En: *Journal of Proteome Research* 8.4 (2009), págs. 1792-1796.

[210] A. I. Nesvizhskii y R. Aebersold. “Interpretation of Shotgun Proteomic Data”. En: *Molecular & Cellular Proteomics* 4.10 (2005), págs. 1419-1440.

[211] A. I. Nesvizhskii, A. Keller, E. Kolker y R. Aebersold. "A statistical model for identifying proteins by tandem mass spectrometry". En: *Analytical Chemistry* 75.17 (2003), págs. 4646-4658.

[212] J. Nickolls, I. Buck, M. Garland y K. Skadron. "Scalable Parallel Programming with CUDA". En: *Queue* 6.2 (mar. de 2008), págs. 40-53.

[213] R. Nogales-Cadenas, S. Jonic, F. Tama, A. A. Arteni, D. Tabas-Madrid, M. Vázquez, A. Pascual-Montano y C. O. Sorzano. "3DEM Loupe: Analysis of macromolecular dynamics using structures from electron microscopy." En: *Nucleic acids research* 41.Web Server issue (2013), W363-W367.

[214] R. Nogales-Cadenas, P. Carmona-Saez, M. Vazquez, C. Vicente, X. Yang, F. Tirado, J. M. Carazo y A. Pascual-Montano. "GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information." En: *Nucleic acids research* 37.Web Server issue (2009), W317-22.

[215] N. A. O'Leary y col. "Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation". En: *Nucleic Acids Research* 44.D1 (2016), págs. D733-D745.

[216] J. C. Oliveros, M. Franch, D. Tabas-Madrid, D. San-León, L. Montoliu, P. Cubas y F. Pazos. "Breaking-Cas—interactive design of guide RNAs for CRISPR-Cas experiments for ENSEMBL genomes". En: *Nucleic Acids Research* 44.W1 (2016), gkw407.

[217] A. Oshlack y M. Wakefield. "Transcript length bias in RNA-seq data confounds systems biology." En: *Biology direct* 4.1 (2009), pág. 14.

[218] P. Paatero y U. Tapper. "Positive matrix factorization: A nonnegative factor model with optimal utilization of error estimates of data values". En: *Environmetrics* 5.2 (1994), págs. 111-126.

[219] Y.-K. Paik, G. S. Omenn, V. Thongboonkerd, G. Marko-Varga y W. S. Hancock. "Genome-wide proteomics, Chromosome-Centric Human Proteome Project (C-HPP), part II." En: *Journal of proteome research* 13.1 (2014), págs. 1-4.

[220] Y.-K. Paik y col. *The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome*. 2012.

[221] B. Palsson. "In silico biology through omics." En: *Nature biotechnology* 20.7 (2002), págs. 649-650.

[222] Q. Pan, O. Shai, L. J. Lee, B. J. Frey y B. J. Blencowe. "Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing". En: *Nature Genetics* 40.12 (2008), págs. 1413-1415.

[223] G. L. Papadopoulos, M. Reczko, V. A. Simossis, P. Sethupathy y A. G. Hatzigeorgiou. "The database of experimentally supported targets: A functional update of TarBase". En: *Nucleic Acids Research* 37.SUPPL. 1 (2009), págs. D155-8.

[224] A. Pascual-Montano, P. Carmona-Saez, M. Chagoyen, F. Tirado, J. M. Carazo y R. D. Pascual-Marqui. "bioNMF: a versatile tool for non-negative matrix factorization in biology." En: *BMC bioinformatics* 7.1 (2006), pág. 366.

[225] D. Patterson. "Future of computer architecture". En: *Berkeley EECS Annual Research Symposium* (2006).

[226] W. R. Pearson y D. J. Lipman. "Improved tools for biological sequence comparison." En: *Proceedings of the National Academy of Sciences of the United States of America* 85.8 (1988), págs. 2444-8.

[227] W. R. Pearson, T. Wood, Z. Zhang y W. Miller. "Comparison of DNA sequences with protein sequences." En: *Genomics* 46.1 (1997), págs. 24-36.

[228] E. Pennisi. "Semiconductors Inspire New Sequencing Technologies". En: *Science* 327.5970 (2010), pág. 1190.

[229] D. N. Perkins, D. J. C. Pappin, D. M. Creasy y J. S. Cottrell. "Probability-based protein identification by searching sequence databases using mass spectrometry data". En: *Electrophoresis* 20.18 (1999), pág. 3551.

[230] G. Piétu y col. "The genexpress IMAGE knowledge base of the human brain transcriptome: A prototype integrated resource for functional and computational genomics". En: *Genome Research* 9.2 (1999), págs. 195-209.

[231] E. A. Ponomarenko y col. "Chromosome 18 transcriptoproteome of liver tissue and HepG2 Cells and targeted proteome mapping in depleted plasma: Update 2013". En: *Journal of Proteome Research* 13.1 (2014), págs. 183-190.

[232] S. Prabakaran, G. Lippens, H. Steen y J. Gunawardena. "Post-translational modification: Nature's escape from genetic imprisonment and the basis for dynamic information encoding". En: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 4.6 (2012), págs. 565-583.

[233] G. Prieto, K. Aloria, N. Osinalde, A. Fullaondo, J. M. Arizmendi y R. Matthiesen. "PAnalyzer: a software tool for protein inference in shotgun proteomics." En: *BMC bioinformatics* 13 (2012), pág. 288.

[234] D. Ramsköld, E. T. Wang, C. B. Burge y R. Sandberg. "An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data". En: *PLoS Computational Biology* 5.12 (2009). Ed. por L. J. Jensen, e1000598.

[235] J. Rappsilber y M. Mann. "What does it mean to identify a protein in proteomics?" En: *Trends in Biochemical Sciences* 27.2 (2002), págs. 74-78.

[236] E Rasmussen. *Clustering algorithms*. Wiley, 1992, pág. 351.

[237] M. Rehmsmeier, P. Steffen, M. Hochsmann y R. Giegerich. "Fast and effective prediction of microRNA/target duplexes." En: *RNA (New York, N.Y.)* 10.10 (2004), págs. 1507-17.

[238] P. H. Reyes-Herrera y E. Ficarra. "One decade of development and evolution of microRNA target prediction algorithms." En: *Genomics, proteomics & bioinformatics* 10.5 (2012), págs. 254-63.

[239] D. R. Rhodes, J. Yu, K Shanker, N. Deshpande, R. Varambally, D. Ghosh, T. Barrette, A. Pandey y A. M. Chinnaiyan. "ONCOMINE: a cancer microarray database and integrated data-mining platform." En: *Neoplasia (New York, N.Y.)* 6.1 (2004), págs. 1-6.

[240] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi y G. K. Smyth. "limma powers differential expression analyses for RNA-sequencing and microarray studies". En: *Nucleic Acids Research* (2015).

[241] M. D. Robinson y G. K. Smyth. "Moderated statistical tests for assessing differences in tag abundance". En: *Bioinformatics* 23.21 (2007), págs. 2881-2887.

[242] M. Ronaghi y col. "A sequencing method based on real-time pyrophosphate." En: *Science (New York, N.Y.)* 281.5375 (1998), págs. 363, 365.

[243] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén y P. L. Nyrén. "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release". En: *Analytical Biochemistry* 242.1 (1996), págs. 84-89.

[244] M. Rosikiewicz, A. Comte, A. Niknejad, R. R. Marc y F. B. Bastian. "Uncovering hidden duplicated content in public transcriptomics data". En: *Database* 2013 (2013), bat010.

[245] R. B. Roth, P. Hevezi, J. Lee, D. Willhite, S. M. Lechner, A. C. Foster y A. Zlotnik. "Gene expression analyses reveal molecular relationships among 20 regions of the human CNS." En: *Neurogenetics* 7.2 (2006), págs. 67-80.

[246] E. Rucci, C. Garcia, G. Botella, A. De Giusti, M. Naiouf y M. Prieto-Matías. "An energy-aware performance analysis of SWIMM: Smith–Waterman implementation on Intel's Multicore and Manycore architectures". En: *Concurrency and Computation: Practice and Experience* 27.18 (2015), págs. 5517-5537.

[247] E. Rucci, C. Garcia, G. Botella, A. De Giusti, M. Naiouf y M. Prieto-Matías. "OSWALD: OpenCL Smith-Waterman Algorithm on Altera FPGA for Large Protein Databases". En: *International Journal of High Performance Computing Applications* (jun. de 2016), pág. 1094342016654215.

[248] J. Rung y A. Brazma. "Reuse of public genome-wide gene expression data". En: *Nature Reviews Genetics* 14.2 (2012), págs. 1-11.

[249] F. Sanger y A. R. Coulson. "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase". En: *Journal of Molecular Biology* 94.3 (1975), págs. 441-448.

[250] S. H. W. Scheres. "RELION: Implementation of a Bayesian approach to cryo-EM structure determination". En: *Journal of Structural Biology* 180.3 (2012), págs. 519-530.

[251] R. Schmieder y R. Edwards. "Quality control and preprocessing of metagenomic datasets". En: *Bioinformatics* 27.6 (2011), págs. 863-864.

[252] R. Schmieder, Y. W. Lim, F. Rohwer y R. Edwards. "TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets." En: *BMC bioinformatics* 11 (2010), pág. 341.

[253] N. J. Schork. “Personalized medicine: Time for one-person trials”. En: *Nature* 520.7549 (2015), págs. 609-611.

[254] J. C. Schwartz, M. W. Senko y J. E. P. Syka. “A two-dimensional quadrupole ion trap mass spectrometer”. En: *Journal of the American Society for Mass Spectrometry* 13.6 (2002), págs. 659-669.

[255] V. Segura y col. “Spanish human proteome project: Dissection of chromosome 16”. En: *Journal of Proteome Research* 12.1 (2013), págs. 112-122.

[256] V. Segura y col. “Surfing transcriptomic landscapes. A step beyond the annotation of chromosome 16 proteome”. En: *Journal of Proteome Research* 13.1 (2014), págs. 158-172.

[257] R. F. Service. “Gene sequencing: The race for the $1000 genome”. En: *Science* 311.5767 (2006), págs. 1544-1546.

[258] P. Sethupathy, B. Corda y A. G. Hatzigeorgiou. “TarBase: A comprehensive database of experimentally supported animal microRNA targets.” En: *RNA (New York, N.Y.)* 12.2 (2006), págs. 192-7.

[259] D. Shalon, S. J. Smith y P. O. Brown. “A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.” En: *Genome Research* 6.7 (1996), págs. 639-645.

[260] M. H. Shen, P. S. Harper y M. Upadhyaya. “Molecular genetics of neurofibromatosis type 1 (NF1).” En: *Journal of medical genetics* 33.1 (1996), págs. 2-17.

[261] G. M. Sheynkman, M. R. Shortreed, B. L. Frey, M. Scalf y L. M. Smith. “Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences.” En: *Journal of proteome research* 13.1 (2014), págs. 228-40.

[262] A. K. Shukla y col. “Protein sequencing by tandem mass spectrometry”. En: *Proc. Natl. Acad. Sci .USA* 83.9 (1986), págs. 6233-6237.

[263] D. Smedley y col. “The BioMart community portal: An innovative alternative to large, centralized data repositories”. En: *Nucleic Acids Research* 43.W1 (2015), W589-W598.

[264] M. Smid y L. C. J. Dorssers. “GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms.” En: *Bioinformatics (Oxford, England)* 20.16 (2004), págs. 2618-25.

[265] G. K. Smith. “limma: Linear Models for Microarray Data”. En: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. 2005. New York: Springer-Verlag, 2005, págs. 397-420.

[266] M. D. Smith. *Support for Speculative Execution in High-Performance Processors*. 1992.

[267] M. Snir. *MPI–the Complete Reference: The MPI core*. 1998, pág. 426.

[268] A. Sodani. “Knights landing (KNL): 2nd Generation Intel® Xeon Phi processor”. En: *Hot Chips 27 Symposium (HCS), 2015 IEEE*. IEEE. 2015, págs. 1-24.

[269] M. S. Son y R. K. Taylor. "Preparing DNA libraries for multiplexed paired-end deep sequencing for Illumina GA sequencers". En: *Current Protocols in Microbiology* Chapter 1.SUPPL.20 (2011), Unit 1E.4.

[270] F. C. Stingo, Y. A. Chen, M. Vannucci, M. Barrier y P. E. Mirkes. "A Bayesian graphical modeling approach to microrna regulatory network inference." En: *The annals of applied statistics* 4.4 (2010), págs. 2024-2048.

[271] K. Struhl. *Fundamentally different logic of gene regulation in eukaryotes and prokaryotes*. 1999.

[272] M. Sturm, M. Hackenberg, D. Langenberger y D. Frishman. "TargetSpy: a supervised machine learning approach for microRNA target prediction." En: *BMC bioinformatics* 11 (2010), pág. 292.

[273] A. Subramanian y col. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." En: *Proceedings of the National Academy of Sciences of the United States of America* 102.43 (2005), págs. 15545-50.

[274] J. Sugerman y G. Venkitachalam. "Virtualizing I/O Devices on VMware Workstation's Hosted Virtual Machine Monitor." En: *USENIX Annual Technical* (2001).

[275] M. Sultan. "A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome". En: *Science* 321.5891 (2008), págs. 956-960.

[276] D. Sun, R. Haddad, J. M. Kraniak, S. D. Horne y M. A. Tainsky. "RAS/MEK–Independent Gene Expression Reveals BMP2-Related Malignant Phenotypes in the Nf1-Deficient MPNST". En: *Molecular Cancer Research* 11.6 (2013), págs. 616-627.

[277] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander y T. R. Golub. "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation." En: *Proceedings of the National Academy of Sciences of the United States of America* 96.6 (1999), págs. 2907-2912.

[278] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees y S. J. Ludtke. "EMAN2: An extensible image processing suite for electron microscopy". En: *Journal of Structural Biology* 157.1 (2007), págs. 38-46.

[279] L. S. Technologies. "Transcriptomics today: Microarrays, RNA-seq, and more". En: *Science* 349.6247 (2015).

[280] R. Thadani y M. T. Tammi. "MicroTar: predicting microRNA targets from RNA duplexes." En: *BMC bioinformatics* 7 Suppl 5 (2006), S20.

[281] The Global Proteome Machine. *cRAP protein sequences*. 2011.

[282] J. F. Thompson y K. E. Steinmann. *Single molecule sequencing with a HeliScope genetic analysis system*. 2010.

[283] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn y L. Pachter. "Differential analysis of gene regulation at transcript resolution with RNA-seq." En: *Nature biotechnology* 31.1 (2013), págs. 46-53.

[284] C. Trapnell, L. Pachter y S. L. Salzberg. "TopHat: Discovering splice junctions with RNA-Seq". En: *Bioinformatics* 25.9 (2009), págs. 1105-1111.

[285] V. Trevino, F. Falciani y H. A. Barrera-Saldaña. "DNA microarrays: a powerful genomic tool for biomedical and clinical research". En: *Molecular Medicine* 13.9-10 (2007), pág. 1.

[286] D. M. Tullsen, S. J. Eggers y H. M. Levy. "Simultaneous multithreading". En: *Proceedings of the 22nd annual international symposium on Computer architecture - ISCA '95*. Vol. 23. 2. New York, New York, USA: ACM Press, 1995, págs. 392-403.

[287] V. G. Tusher, R. Tibshirani y G. Chu. "Significance analysis of microarrays applied to the ionizing radiation response." En: *Proceedings of the National Academy of Sciences of the United States of America* 98.9 (2001), págs. 5116-21.

[288] M. Tyers, M. Tyers, M. Mann y M. Mann. "From genomics to proteomics." En: *Nature* 422.March (2003), págs. 193-7.

[289] A. Valouev y col. "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning". En: *Genome Research* 18.7 (2008), págs. 1051-1063.

[290] R. N. Van Gelder, M. E. von Zastrow, a Yool, W. C. Dement, J. D. Barchas y J. H. Eberwine. "Amplified RNA synthesized from limited quantities of heterogeneous cDNA." En: *Proceedings of the National Academy of Sciences of the United States of America* 87.5 (1990), págs. 1663-1667.

[291] M. Vazquez, R. Nogales-Cadenas, J. Arroyo, P. Botías, R. García, J. M. Carazo, F. Tirado, A. Pascual-Montano y P. Carmona-Saez. "MARQ: an online tool to mine GEO for experiments with similar or opposite gene expression signatures." En: *Nucleic acids research* 38.Web Server issue (2010), W228-32.

[292] VCFv y BCFv. *The Variant Call Format Specification*. 2015.

[293] S. Velankar, J. M. Dana, J. Jacobsen, G. Van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M. J. Martin y G. J. Kleywegt. "SIFTS: Structure Integration with Function, Taxonomy and Sequences resource". En: *Nucleic Acids Research* 41.D1 (2013), págs. D483-D489.

[294] V. Velculescu, L. Zhang y B. Vogelstein. "Serial analysis of gene expression". En: *Science* 270.5235 (1995), págs. 484-487.

[295] V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett, P. Hieter, B. Vogelstein y K. W. Kinzler. "Characterization of the yeast transcriptome". En: *Cell* 88.2 (1997), págs. 243-251.

[296] P. M. Visscher, M. A. Brown, M. I. McCarthy y J. Yang. "Five years of GWAS discovery". En: *American Journal of Human Genetics* 90.1 (2012), págs. 7-24.

[297] R. Vita y col. "The immune epitope database (IEDB) 3.0". En: *Nucleic Acids Research* 43.D1 (2015), págs. D405-D412.

[298] J. A. Vizcaíno y col. "ProteomeXchange provides globally coordinated proteomics data submission and dissemination." En: *Nature biotechnology* 32.3 (2014), págs. 223-6.

[299] G. P. Wagner, K. Kin y V. J. Lynch. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples". En: *Theory in Biosciences* 131.4 (2012), págs. 281-285.

[300] X. Wang y B. Zhang. "customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search." En: *Bioinformatics (Oxford, England)* 29.24 (2013), págs. 3235-7.

[301] Y. Wang, G. Kora, B. P. Bowen y C. Pan. "MIDAS: A database-searching algorithm for metabolite identification in metabolomics". En: *Analytical Chemistry* 86.19 (2014), págs. 9496-9503.

[302] Z. Wang, M. Gerstein y M. Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." En: *Nature reviews. Genetics* 10.1 (2009), págs. 57-63.

[303] V. C. Wasinger, S. J. Cordwell, A. CerpaPoljak, J. X. Yan, A. A. Gooley, M. R. Wilkins, M. W. Duncan, R. Harris, K. L. Williams e I. Humphery-Smith. "Progress with geneproduct mapping of the Mollicutes: Mycoplasma genitalium". En: *ELECTROPHORESIS* 16.1 (1995), págs. 1090-1094.

[304] E. V. Wasmuth y C. D. Lima. "OUP accepted manuscript". En: *Nucleic Acids Research* 45.D1 (2016), págs. 1-15.

[305] E. V. Wasmuth y C. D. Lima. "UniProt: the universal protein knowledgebase". En: *Nucleic Acids Research* 45.November 2016 (2016), págs. 1-12.

[306] J. Watson. "Virtualbox: bits and bytes masquerading as machines". En: *Linux Journal* 2008.166 (2008), pág. 1.

[307] B. T. Wilhelm y J. R. Landry. *RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing*. 2009.

[308] M. Wilhelm y col. "Mass-spectrometry-based draft of the human proteome." En: *Nature* 509.7502 (2014), págs. 582-7.

[309] P. Wirapati y col. "Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures". En: *Breast Cancer Research* 10.4 (2008), R65.

[310] T. J. Wu, A. Shamsaddini, Y. Pan, K. Smith, D. J. Crichton, V. Simonyan y R. Mazumder. "A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual Environment (HIVE)". En: *Database* 2014.0 (2014), bau022.

[311] W. A. Wulf y S. A. McKee. "Hitting the memory wall: Implications of the Obvious". En: *ACM SIGARCH Computer Architecture News* 23.1 (1995), págs. 20-24.

[312] J. Xin y col. "High-performance web services for querying gene and variant annotation". En: *Genome Biology* 17.1 (2016), pág. 91.

[313] C. Yamasaki y col. “The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts.” En: *Nucleic acids research* 36.Database issue (2008), págs. D793-9.

[314] J.-H. Yang, J.-H. Li, P. Shao, H. Zhou, Y.-Q. Chen y L.-H. Qu. “starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data.” En: *Nucleic acids research* 39.Database issue (2011), págs. D202-9.

[315] A. Yates y col. “Ensembl 2016”. En: *Nucleic Acids Research* 44.D1 (2016), págs. D710-D716.

[316] M. Yousef, S. Jung, A. V. Kossenkov, L. C. Showe y M. K. Showe. “Naïve Bayes for microRNA target predictions–machine learning for microRNA targets.” En: *Bioinformatics (Oxford, England)* 23.22 (2007), págs. 2987-92.

[317] D. Yue, M. Guo, Y. Chen e Y. Huang. “A Bayesian decision fusion approach for microRNA target prediction.” En: *BMC genomics* 13 Suppl 8.Suppl 8 (2012), S13.

[318] Y. Zhang, B. R. Fonslow, B. Shan, M. C. Baek y J. R. Yates. “Protein analysis by shotgun/bottom-up proteomics”. En: *Chemical Reviews* 113.4 (2013), págs. 2343-2394.

[319] S. Zhao y W. B. Bruce. “Expression profiling using cDNA microarrays.” En: *Methods in molecular biology (Clifton, N.J.)* 236.1 (2003), págs. 365-80.